A potpourri of analysis considerations for microbial community function

Eric Franzosa

Kelsey Thompson

Curtis Huttenhower (chuttenh@hsph.harvard.edu)

2022-07-27

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

THE HARVARD CHAN MICROBIOME IN PUBLIC HEALTH CENTER

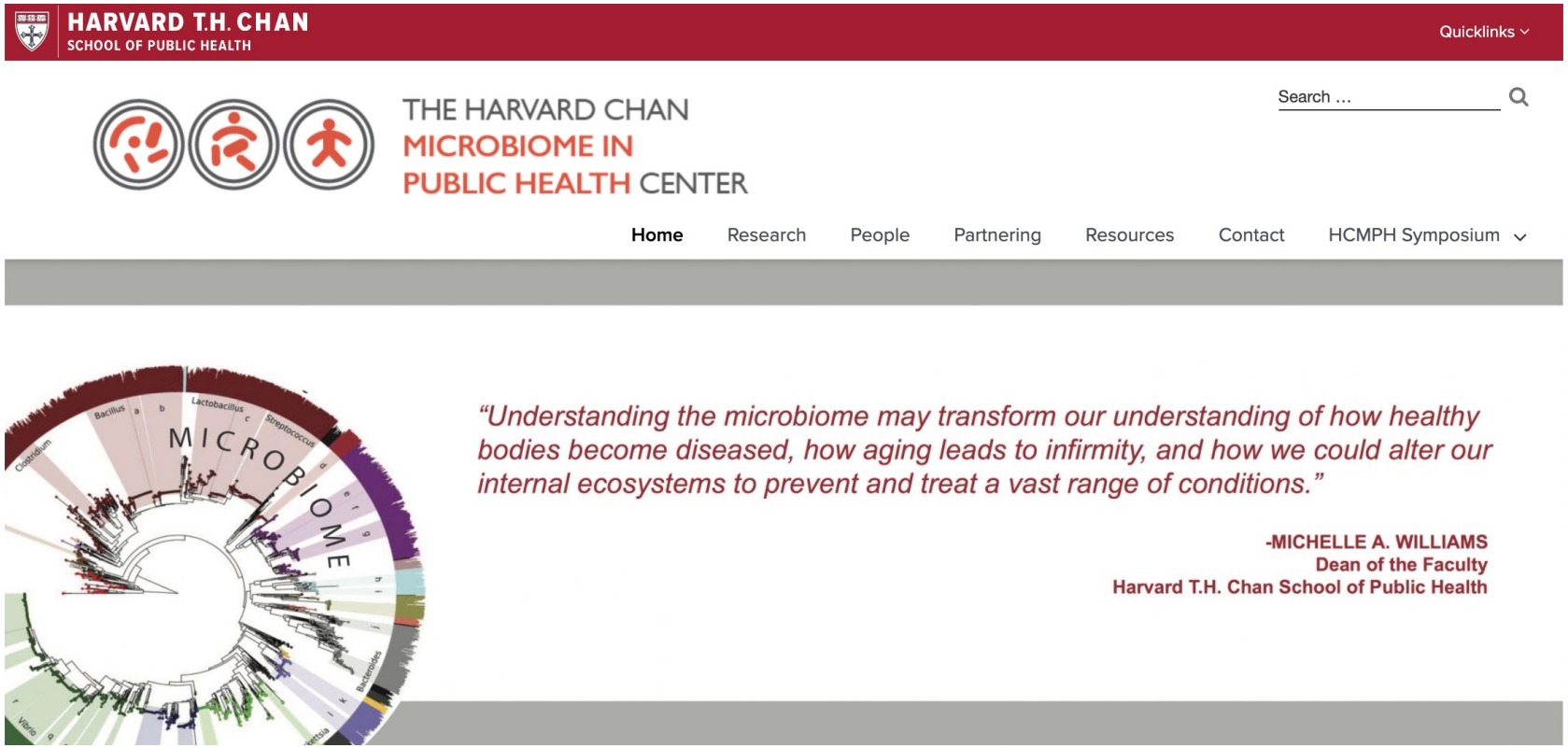
BROAD INSTITUTE

Who am I and why am I talking to you?

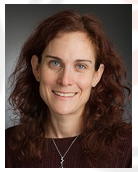
- Background originally computer science and chemistry.
- Worked in the software industry for a few years.
- Got into computational biology during grad school.
- PhD in computational functional genomics / microbiology.
- Worked on microbial communities at Harvard since 2009 (!)
- Because they're cool!
  And because they matter for population health.

The mission of the HCMPH

Use an understanding of the microbiome to improve population health, through basic research, translation, policy, education, and outreach.



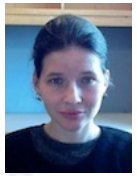

Curtis Huttenhower


Wendy Garrett


Andy Chan


Eric Rimm

https://hcmph.sph.harvard.edu


Xochitl Morgan

HCMPH activities

- **Research**
  - Large flagship consortium projects: OPTIMISTICC, HMBR, HMP, Micro-N…
  - Microbiome population and basic science studies

- **Facilities**
  - BIOM-Mass: BIObank for Microbiome research in Massachusetts *(and elsewhere)*
  - Harvard Chan Microbiome Collection and Analysis Cores
  - Harvard Chan Gnotobiotics Facility

- **Training**
  - IID209: The Human Microbiome and Microbial Communities
  - Year-round methodological short-courses and boot camps
    - Harvard Catalyst Introduction to ‘Omics, Biomarker Science, Network Medicine

- **Community**
  - Microbiome Epidemiology Working Group (MEWG)
  - HCMPH monthly Scientific Meeting, annual Symposium

- **Collaboration**
  - Harvard Chan School, Longwood, Broad Institute, national, international…
  - Industry partnerships

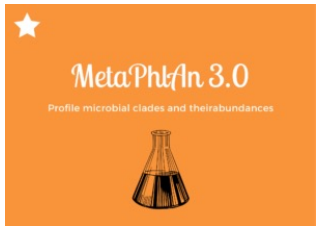
Stuff we do in the lab

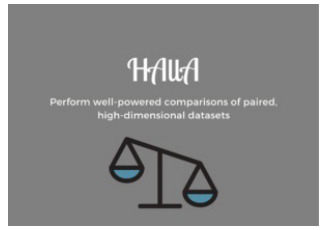
- About 85% computational:
  - ~Half methods development for microbial communities generally.
    - Computational
    - Statistics
  - ~Half applications.
    - Mostly human microbiome epidemiology.
    - Often gut, often inflammatory disease (IBD, arthritis, cancer, etc.)
- The rest is in the wet lab...
  - Data generation
  - Technology development
  - Validation experiments
- *How do microbial communities work and why?*

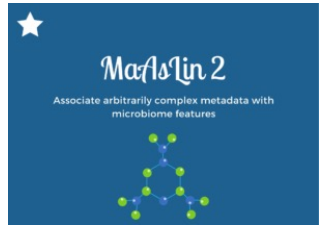
The bioBakery: a next-generation environment for microbiome analyses

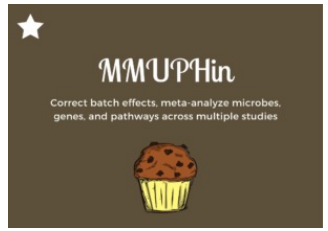
https://huttenhower.sph.harvard.edu/biobakery

- Environment for meta'ome analysis
  - Shotgun metagenomes/metatranscriptomes, amplicons (16S/ITS)
  - Taxonomic and functional profiling
  - Experimental design, statistical analysis
- Pre-built one-click environment to run:
  - On your laptop graphically
  - On a server remotely
  - On the cloud









Microbial community

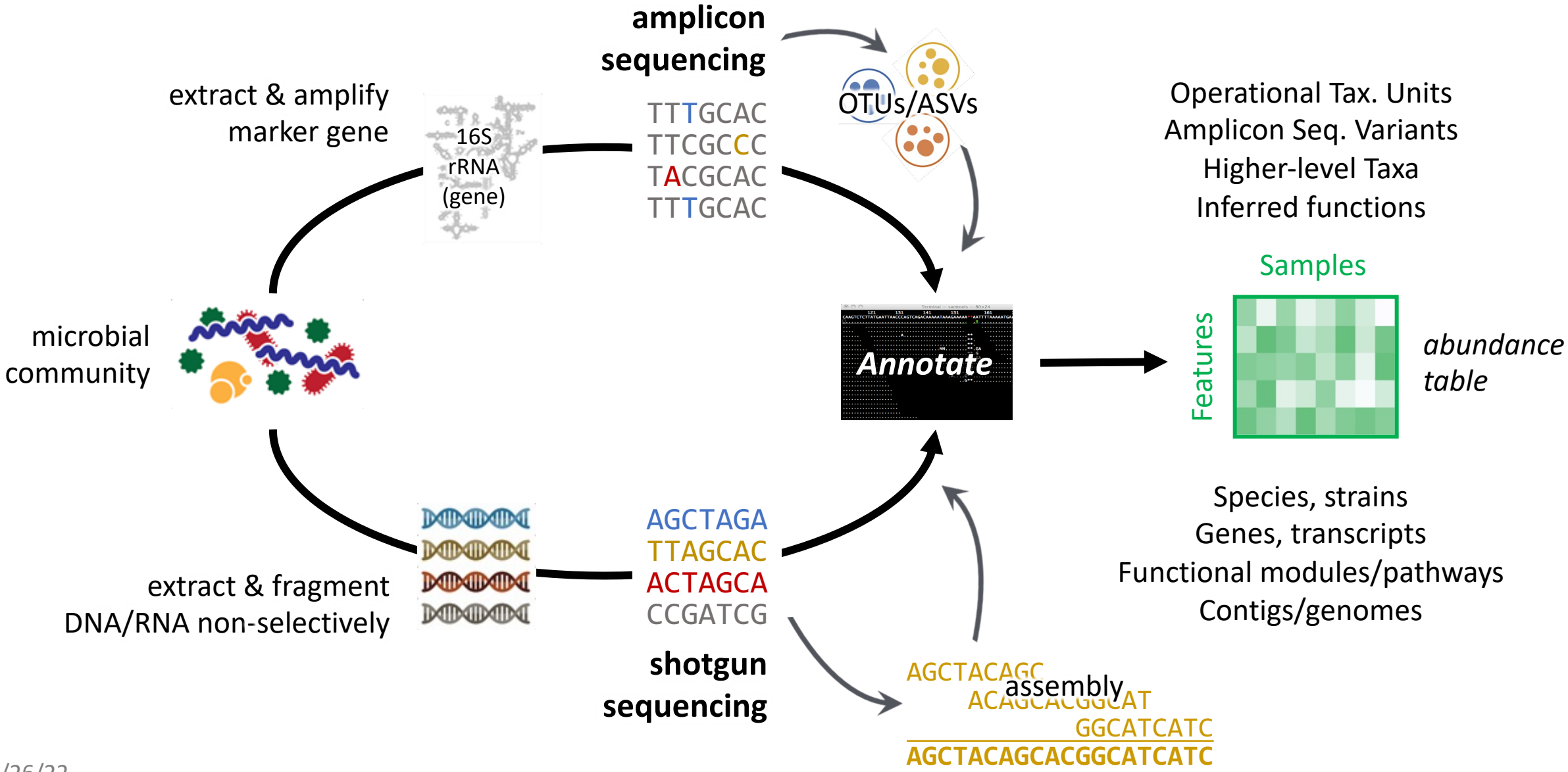
(metagenomic)

functional profiling

Summary

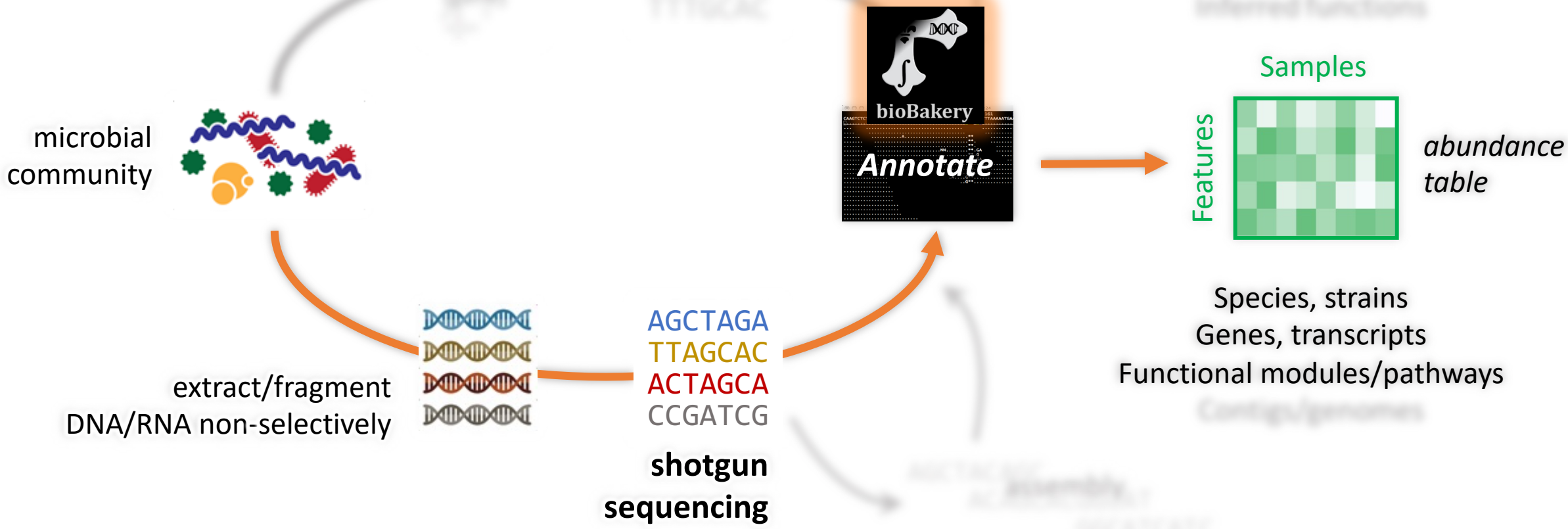
- Introduction to functional profiling
- Tiered search with HUMAnN
- Core functions of the human microbiome
- Contributional diversity
- What's new in bioBakery 3.0?
- Prioritizing bioactive gene families with MetaWIBELE

Everything I learned so far at STAMPS in one slide



bioBakery 3 directly annotates meta'omic reads

- bioBakery 1 c. 2012
- bioBakery 2 c. 2015 - 2018
- **bioBakery 3 c. 2020**

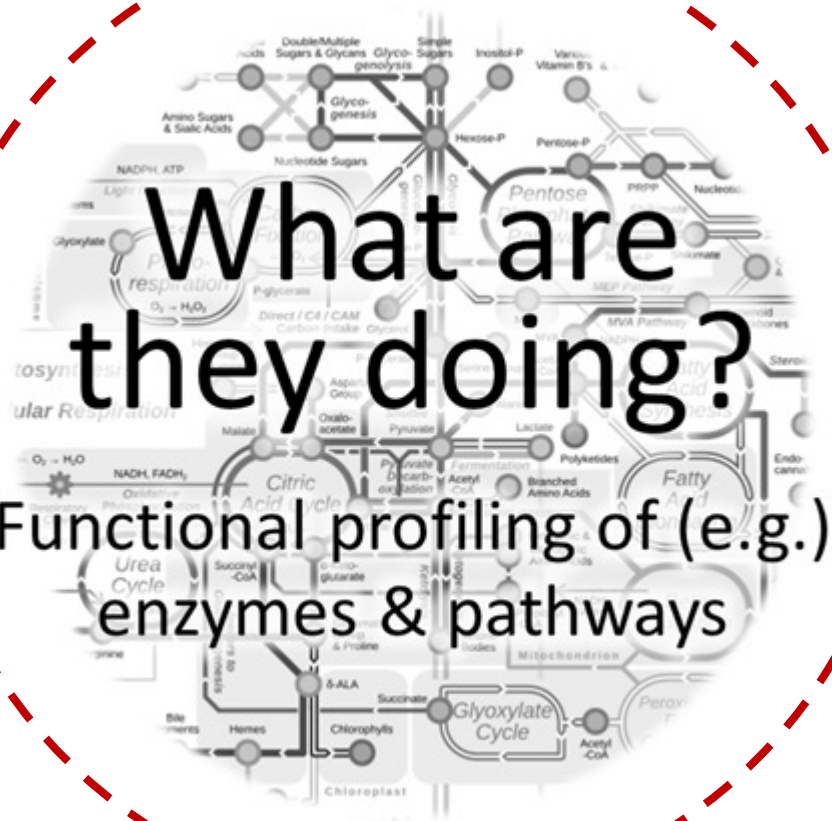


http://huttenhower.sph.harvard.edu/biobakery

Two big questions of shotgun meta'omics

Who is there?

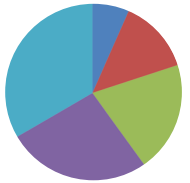
Taxonomic profiling of (e.g.)
phyla, species & strains



What are they doing?

Functional profiling of (e.g.)
enzymes & pathways

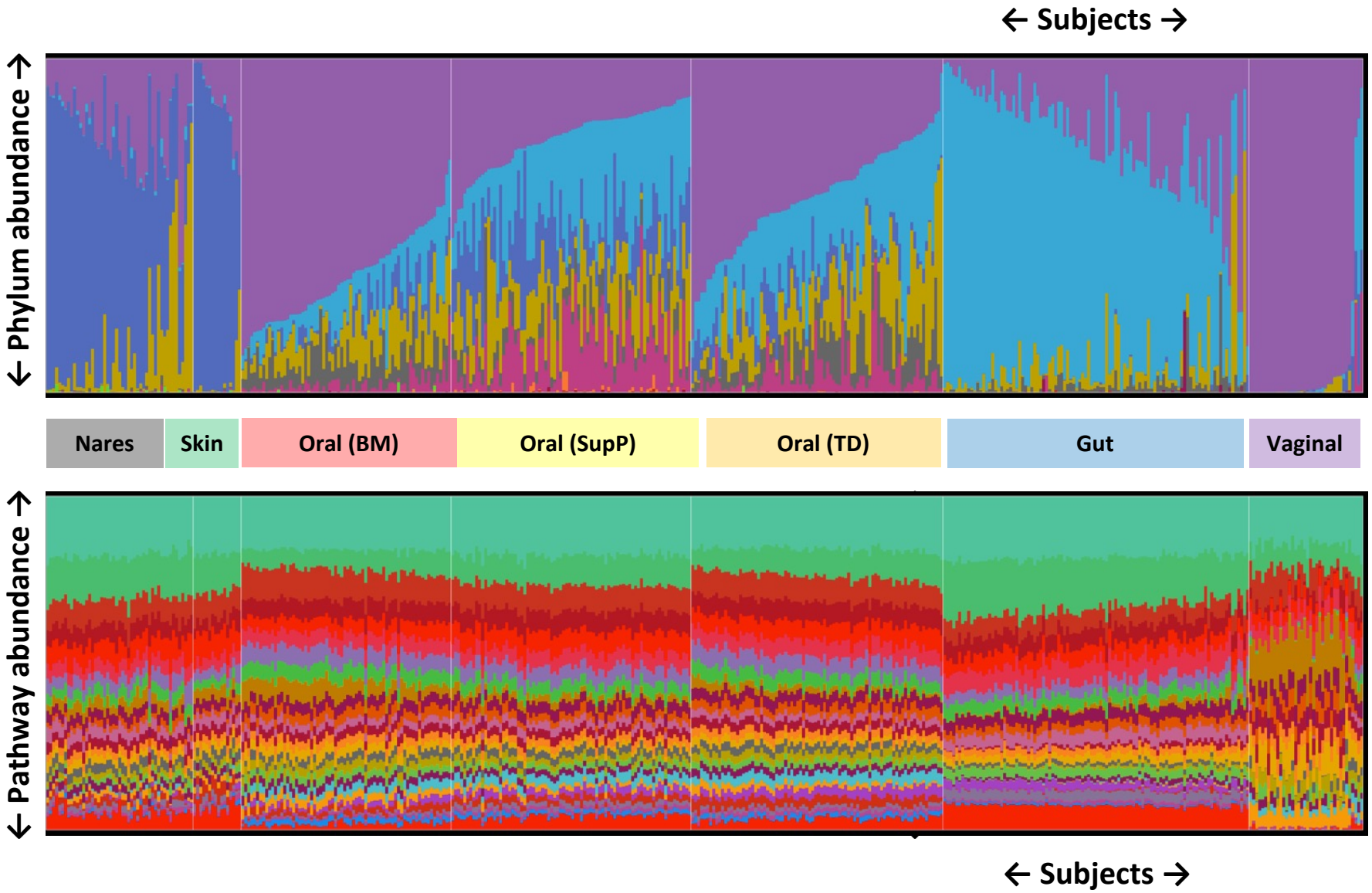
What is a functional profile?



- A functional profile is a vector of relative abundance values describing the molecular functions of a microbial community biosample
- The abundances correspond to features at a fixed level of resolution
  - Millions of conserved gene families (e.g. UniRef90)
  - 10Ks of broad gene families / enzymatic functions (e.g. UniRef50, KO, eggNOG)
  - 100s of metabolic pathways (e.g. KEGG, MetaCyc)
- The abundances form a composition (i.e. they sum to 100%)
- Abundances are proportional to function coverage in the sequencing pool
- Abundances are proportional to copy number in the underlying biosample

Why care about microbiome function?

HMP,
Nature, 2012



Community success in an environment requires being able to do specific tasks (functions)

Different “bags of functions” (i.e. taxa) can collectively satisfy requirements

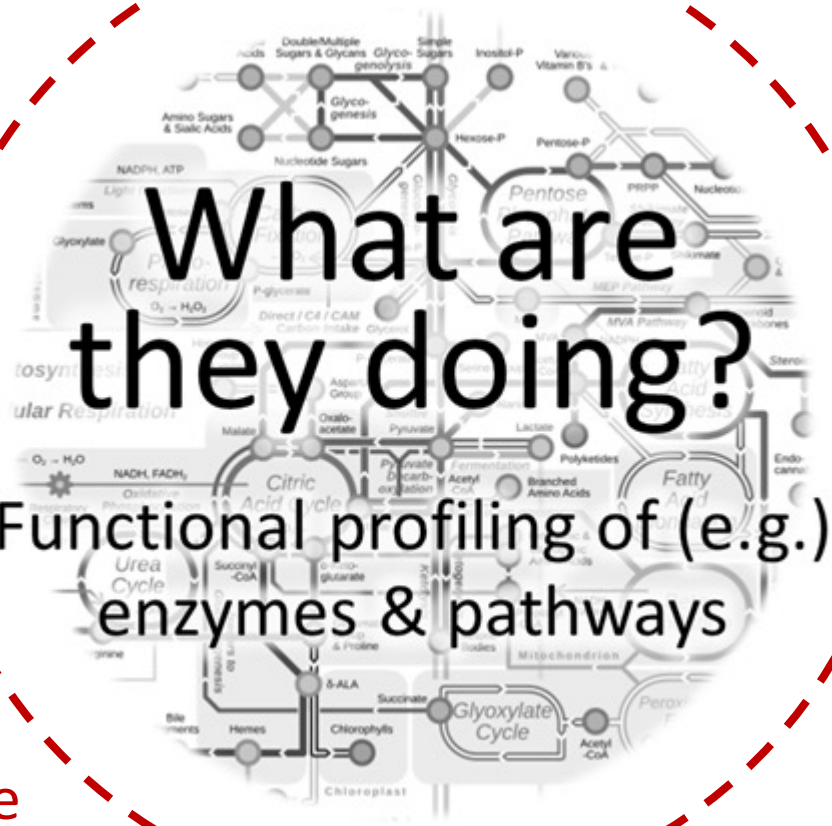
Thus, community function is more conserved than structure/taxonomy

Two big questions of shotgun meta'omics

We often want to answer these at the same time

Who is there?

Taxonomic profiling of (e.g.) phyla, species & strains



What are they doing?

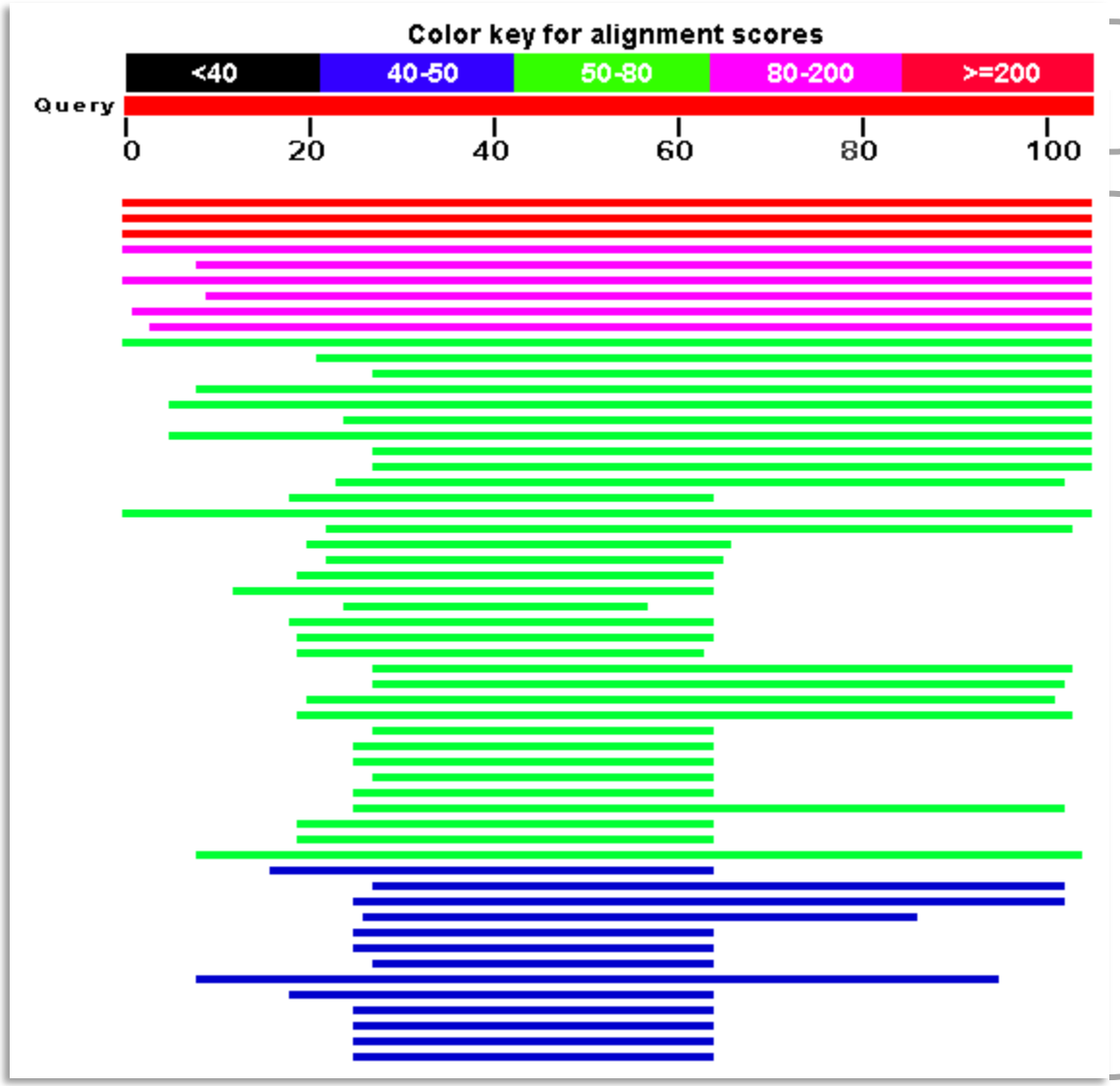
Functional profiling of (e.g.) enzymes & pathways

Which organisms are supplying which functions?

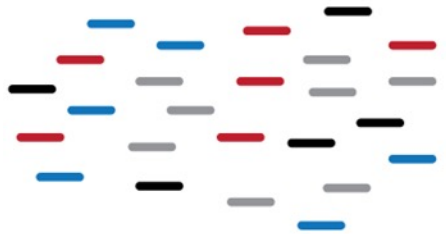
Functional profiling by homology-based search

Profiling meta'omes by homology-based search

Three key ingredients:



1. Queries



A quality-controlled metagenome or metatranscriptome (DNA/RNA reads)

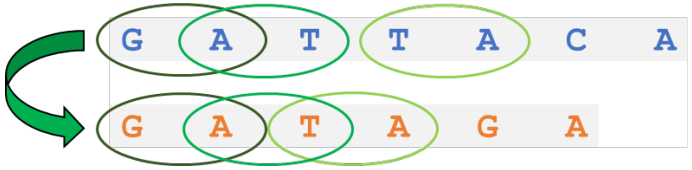
2. Database

Of sequences with which to compare (genomes, genes, proteins)

3. Search "Engine"

Of which BLAST is the prototypical example, though we often prefer other, more specialized methods

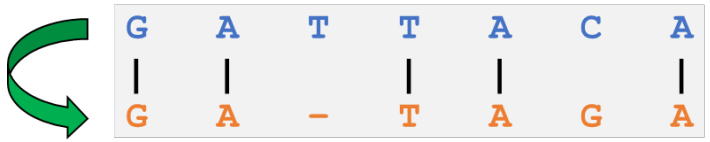
3. Search engines: Alignment and mapping



A *read mapping* places reads into source sequences without using alignment (usually based on overlap of *k*-mer words)

Example Tools

Kallisto, Salmon (general mappers)
Clark, Kraken (taxonomic binners)



A *nucleotide alignment* maps homologous sites between the query and database sequences (inserting gaps to help)

Example Tools

blastn (for fewer, longer sequences)
bowtie2, *bwa-mem* (for short reads)

A *translated alignment* compares the 6 possible translations of a DNA sequence to a protein sequence database

Example Tools

blastp (classic but slow)
DIAMOND, MMSeqs2 (fast)

Faster

More specific (fewer false positives)

Less sensitive (more false negatives)

Trade-offs

Slower

Less specific (more false positives)

More sensitive (fewer false negatives)

3. Search engines: Alignment and mapping

Translated search is a powerful option for functional profiling because it is sensitive to remote homology

However, this sensitivity makes translated search more prone to false positive hits

And, in addition, it is very slow compared to nucleotide-level mapping/alignment (*6 protein searches per nucleotide read*)

A *translated alignment* compares the 6 possible translations of a DNA sequence to a protein sequence database

Example Tools

blastp (classic but slow)

DIAMOND, MMSeqs2 (fast)

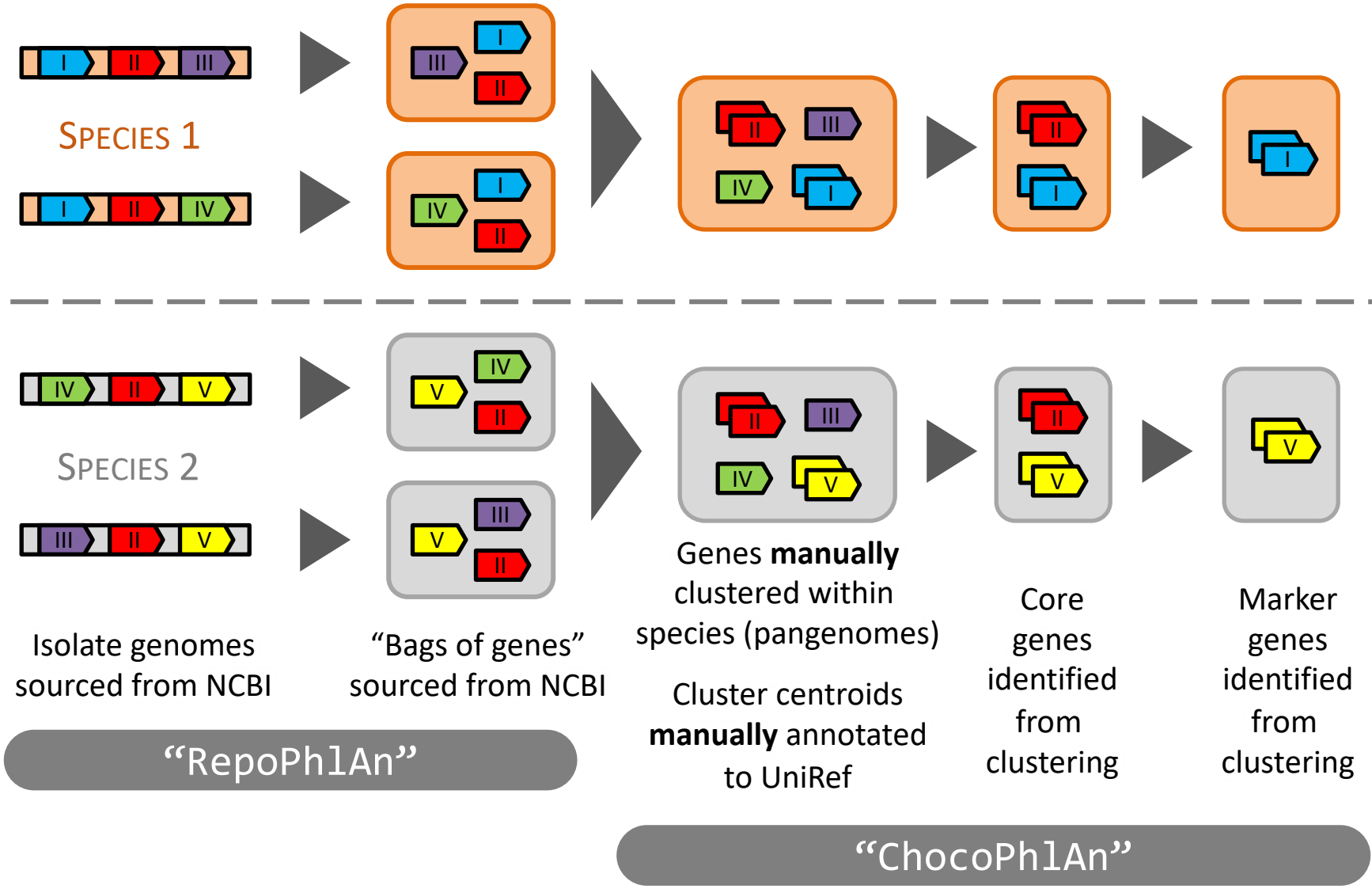
Slower

Less specific (more false positives)

More sensitive (fewer false negatives)

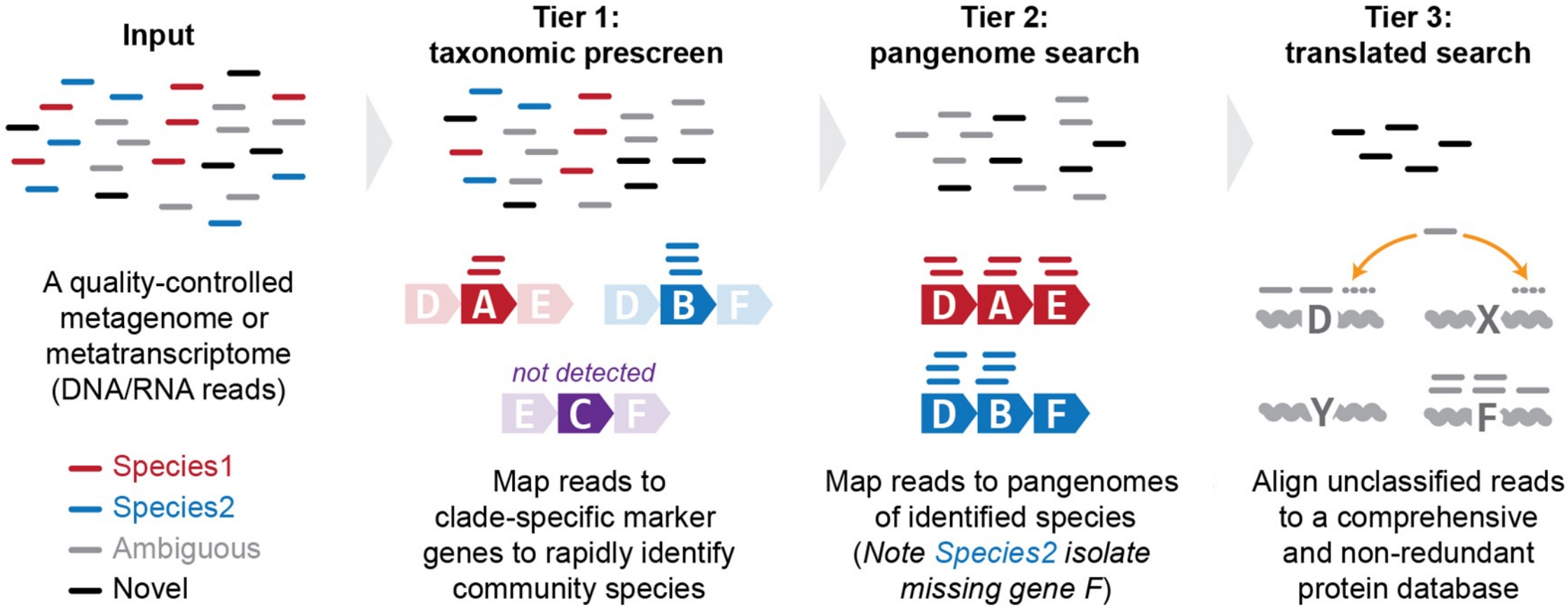
The HUMAnN approach to functional profiling

HUMAnN takes advantage of functionally annotated pangenomes and marker-gene based species detection



- RepoPhlAn and ChocoPhlAn are database systems for identifying markers genes from isolate genomes (among other tasks).
- We will use other products of these systems in subsequent analysis methods.

Accurate, efficient meta'omic profiling by tiered search



See **KneadData** for meta'omic QC

MetaPhlAn **HUMAnN**

http://huttenhower.sph.harvard.edu/humann

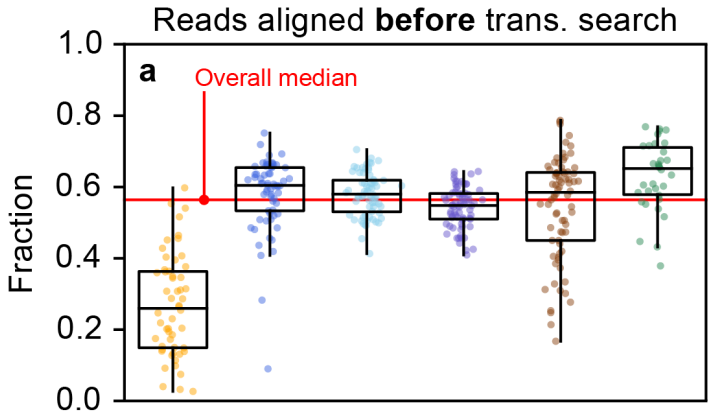
HUMAnN's tiered search is a more accurate approach

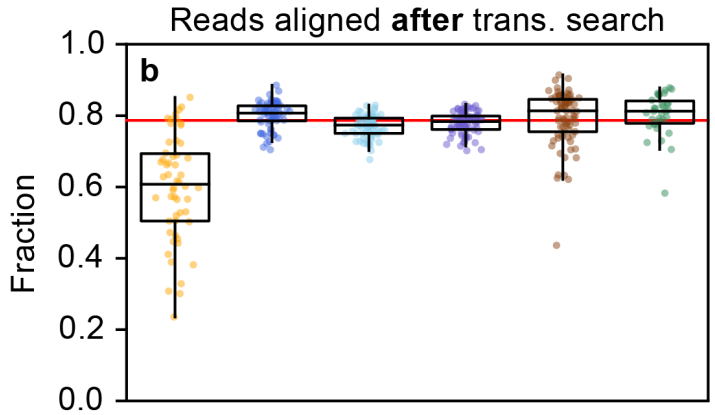
- Evaluation on a synthetic gut metagenome (top-20 HMP)
- 0.1x to 100x coverage
- Compare expected and observed gene coverage
- Pure translated search loses specificity (false positives)
- HUMAnN's tiered search is more accurate overall
- Per-species results for free!
  - Specific for all species
  - sensitive when coverage > 1x

Tiered search is also faster!

If 1/T reads unaligned after pangenome search, then HUMAnN is T times faster than pure translated search (e.g. ½ → 2x faster)

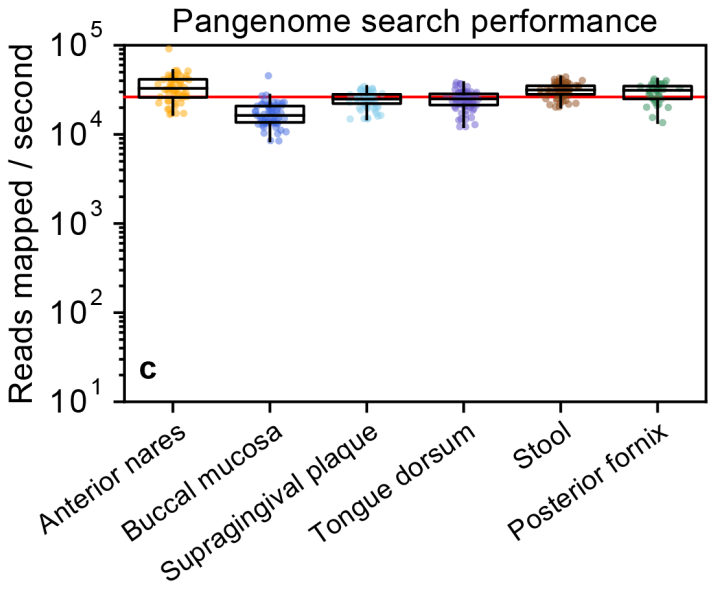
Lots of reads explained during fast search (or this would be a moot point)

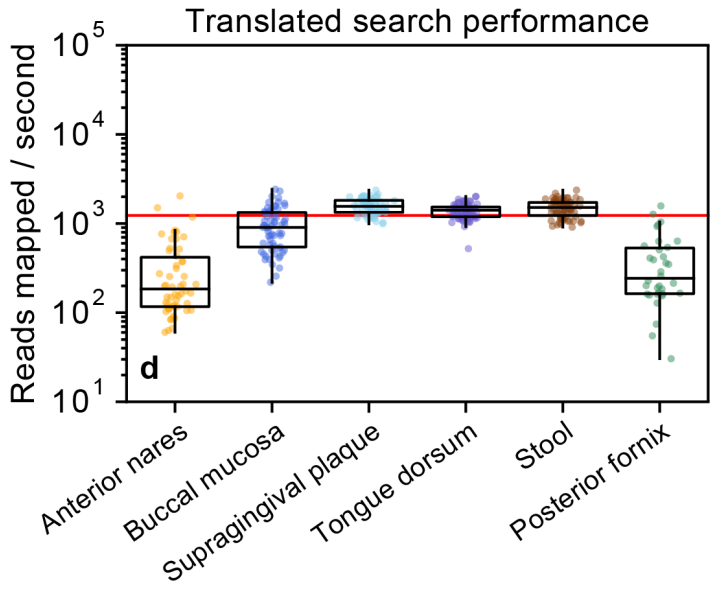




Additional reads align in translated search, especially those from skin

Reads align 10x faster to selected pangenomes vs. translated search





Consequence of a smaller database and nucleotide-level alignment

HUMAnN's species-stratified abundance format

Tiered search is *faster*, *more accurate*, and also provides *more biological information*

UniRef90_R6K3Z5: IMP dehydrogenase	600.95
UniRef90_R6K3Z5: IMP dehydrogenase|Bacteroides_caccae	234.76
UniRef90_R6K3Z5: IMP dehydrogenase|Bacteroides_dorei	107.38
UniRef90_R6K3Z5: IMP dehydrogenase|Bacteroides_ovatus	92.18
UniRef90_R6K3Z5: IMP dehydrogenase|Bacteroides_stercoris	83.95
UniRef90_R6K3Z5: IMP dehydrogenase|Bacteroides_vulgatus	57.27
UniRef90_R6K3Z5: IMP dehydrogenase|unclassified	25.41

HUMAnN's species-stratified abundance format

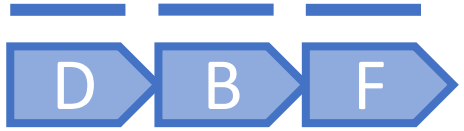
UniRef gene family ID	UniRef gene family name	Community-total gene abundance (RPK)
UniRef90_R6K3Z5:	IMP dehydrogenase	600.95
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_caccae	234.76
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_dorei	107.38
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_ovatus	92.18
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_stercoris	83.95
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_vulgatus	57.27
UniRef90_R6K3Z5:	IMP dehydrogenase|unclassified	25.41

Per-species & unclassified stratifications

Regrouping genes to pathways/modules

- If a species is present, we expect to see its marker genes evenly covered:

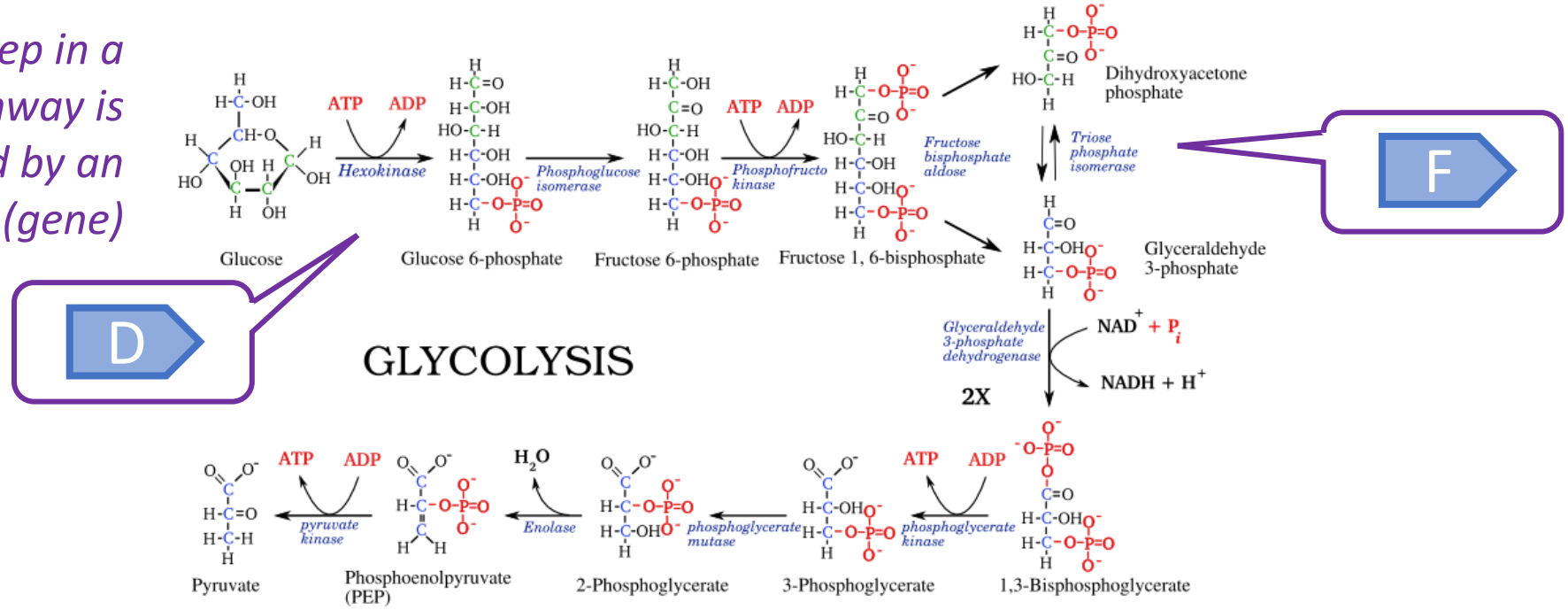
Species 1
(2+3+1)/3 = 2x



Species 2
(1+1+1)/3 = 1x

- If a pathway is present, we expect to see its enzymes evently covered:

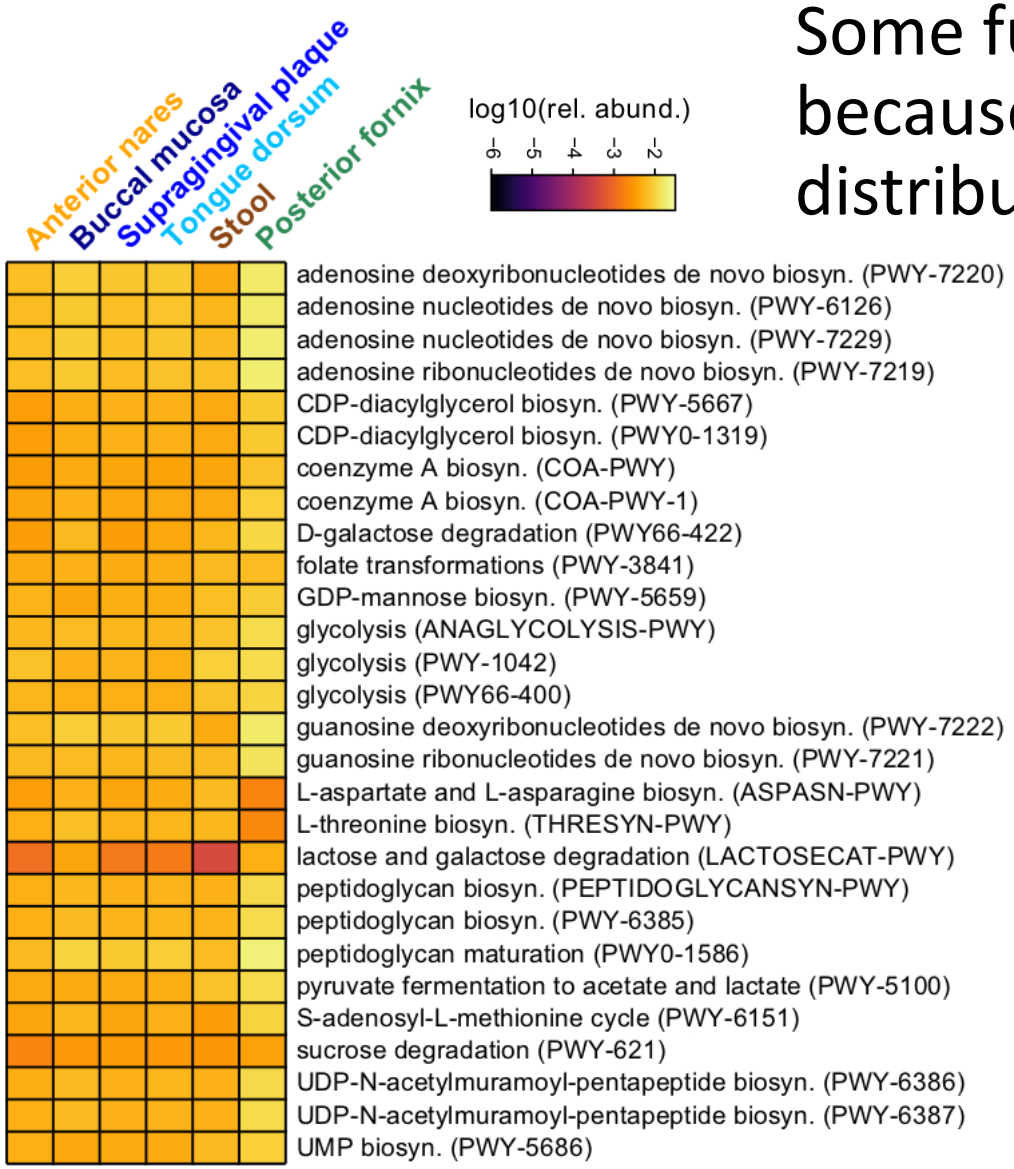
Each step in a metabolic pathway is governed by an enzyme (gene)



Core functions of the human microbiome & contributional diversity

Core functions of the human microbiome

Lloyd-Price, Nature, 2017

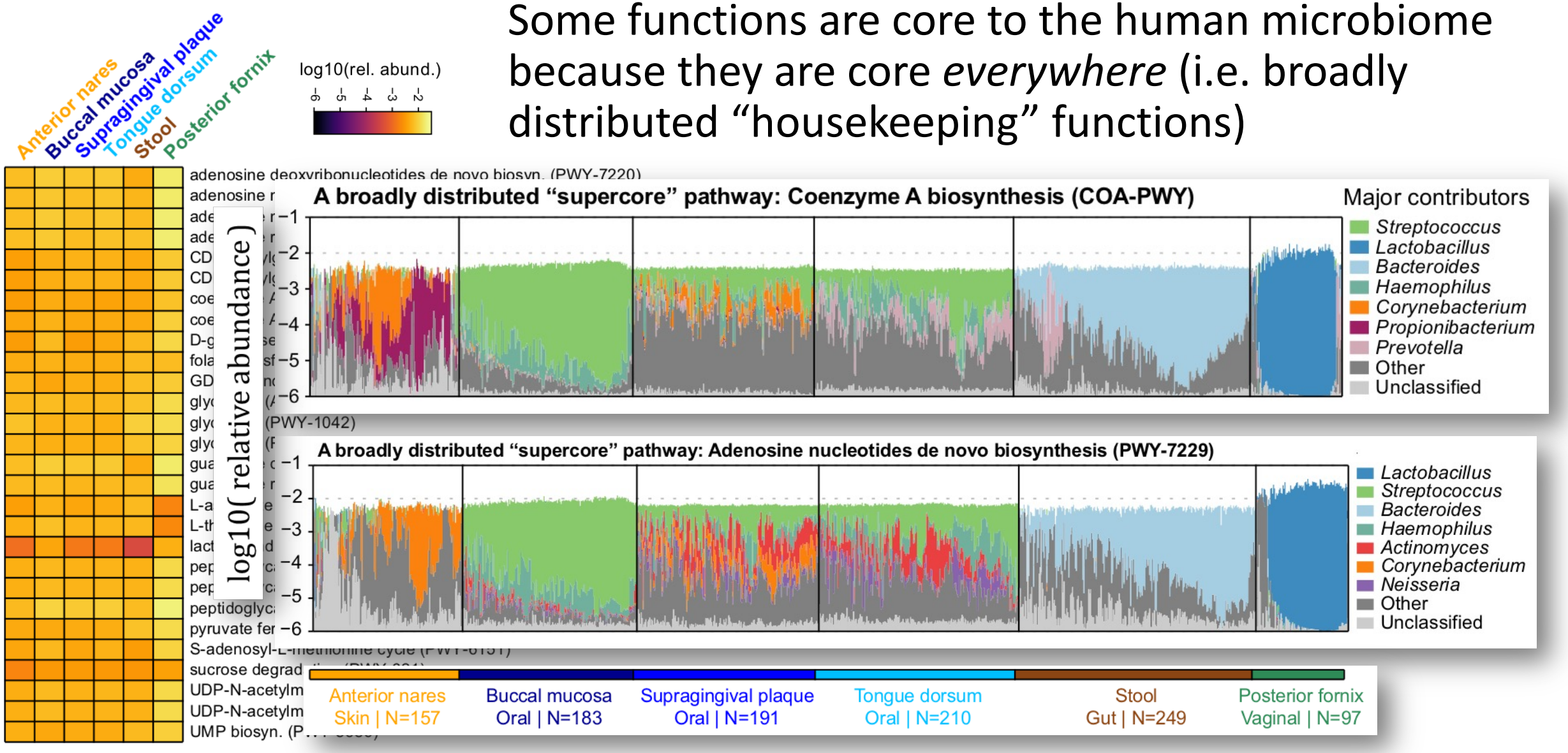


Some functions are core to the human microbiome because they are core *everywhere* (i.e. broadly distributed “housekeeping” functions)

Core functions of the human microbiome

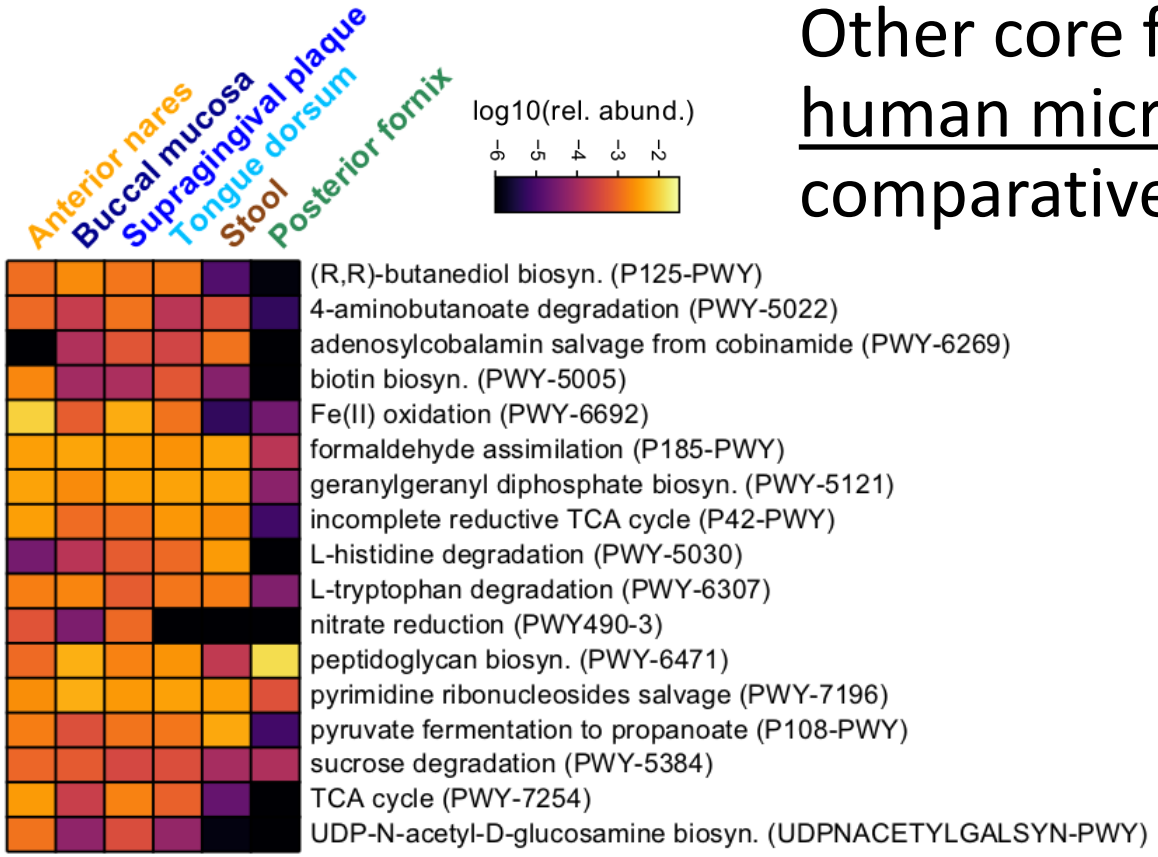
Lloyd-Price, Nature, 2017

Some functions are core to the human microbiome because they are core *everywhere* (i.e. broadly distributed "housekeeping" functions)

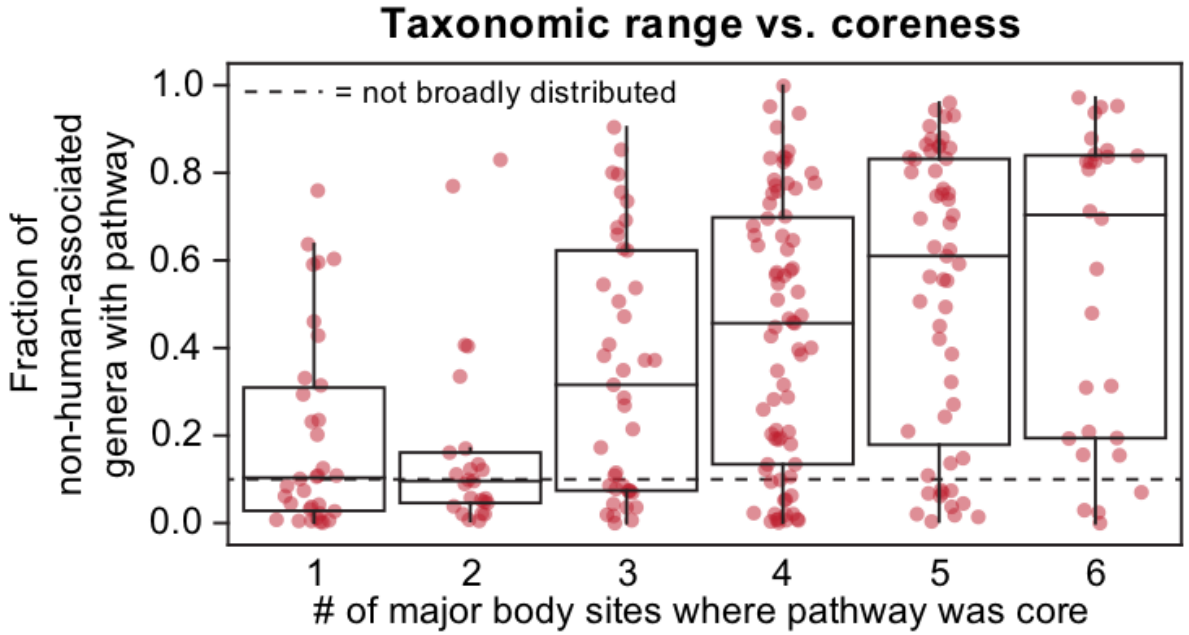


Core functions of the human microbiome

Lloyd-Price, Nature, 2017



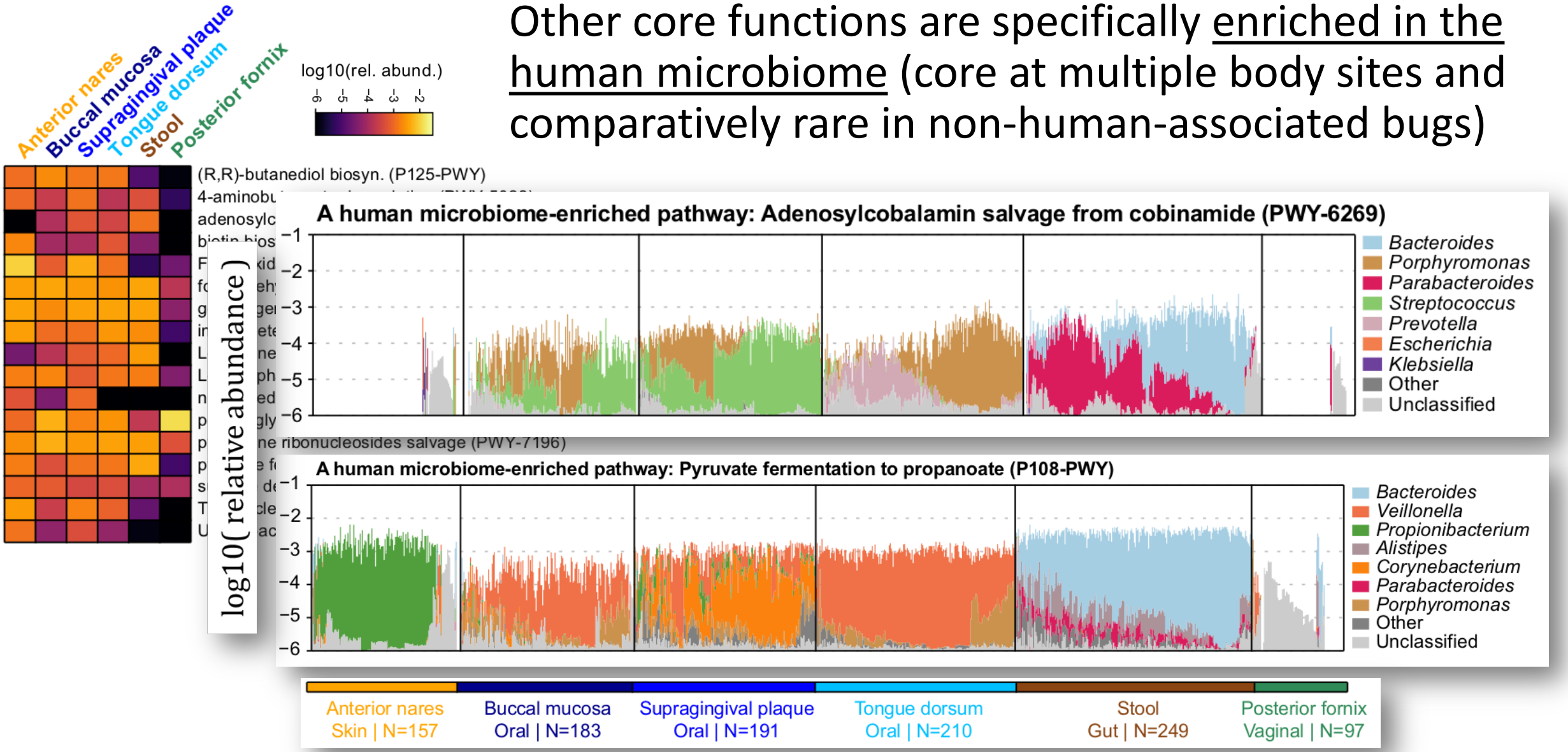
Other core functions are specifically <u>enriched in the human microbiome</u> (core at multiple body sites and comparatively rare in non-human-associated bugs)



Core functions of the human microbiome

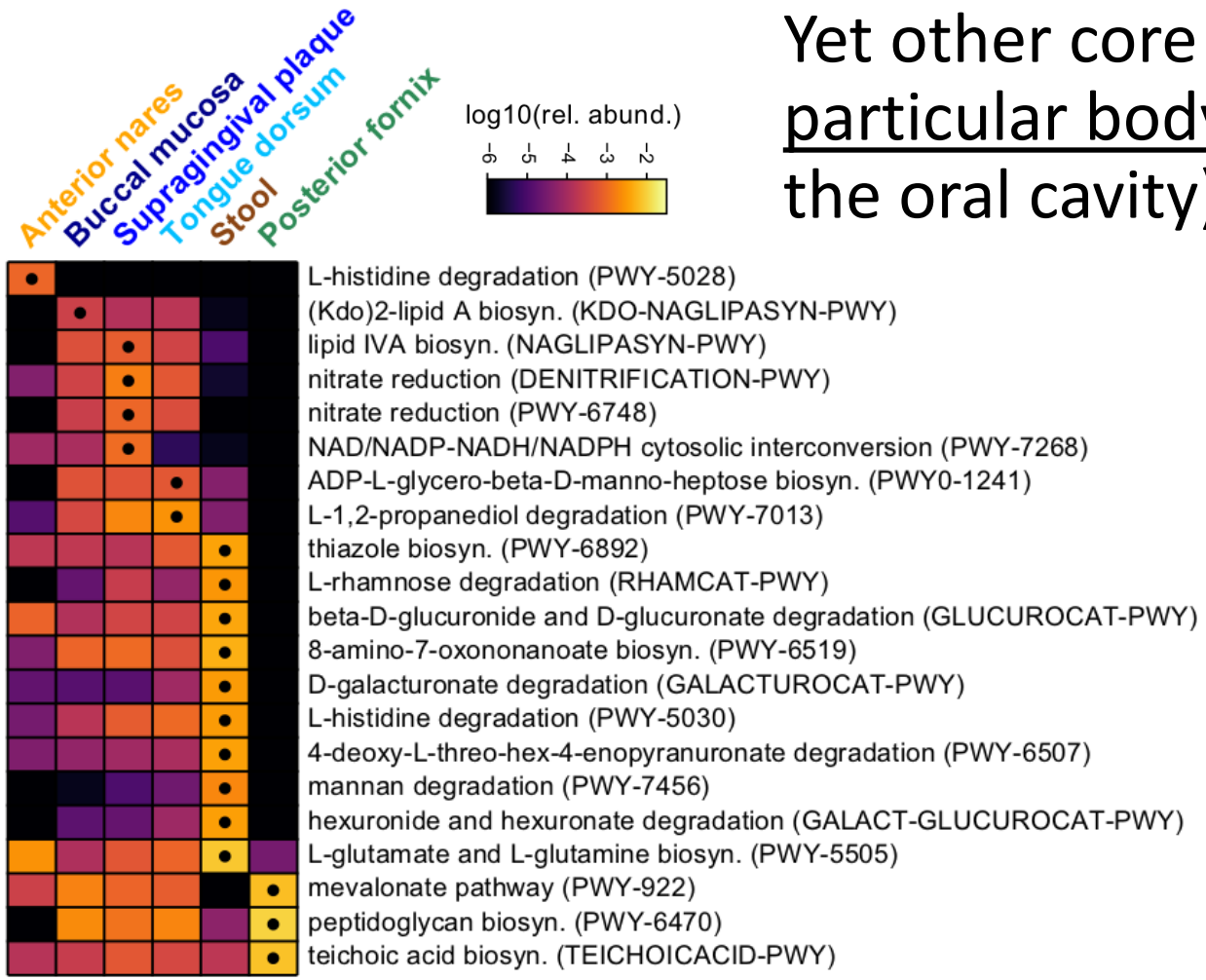
Lloyd-Price, Nature, 2017

Other core functions are specifically <u>enriched in the human microbiome</u> (core at multiple body sites and comparatively rare in non-human-associated bugs)



Core functions of the human microbiome

Lloyd-Price, Nature, 2017

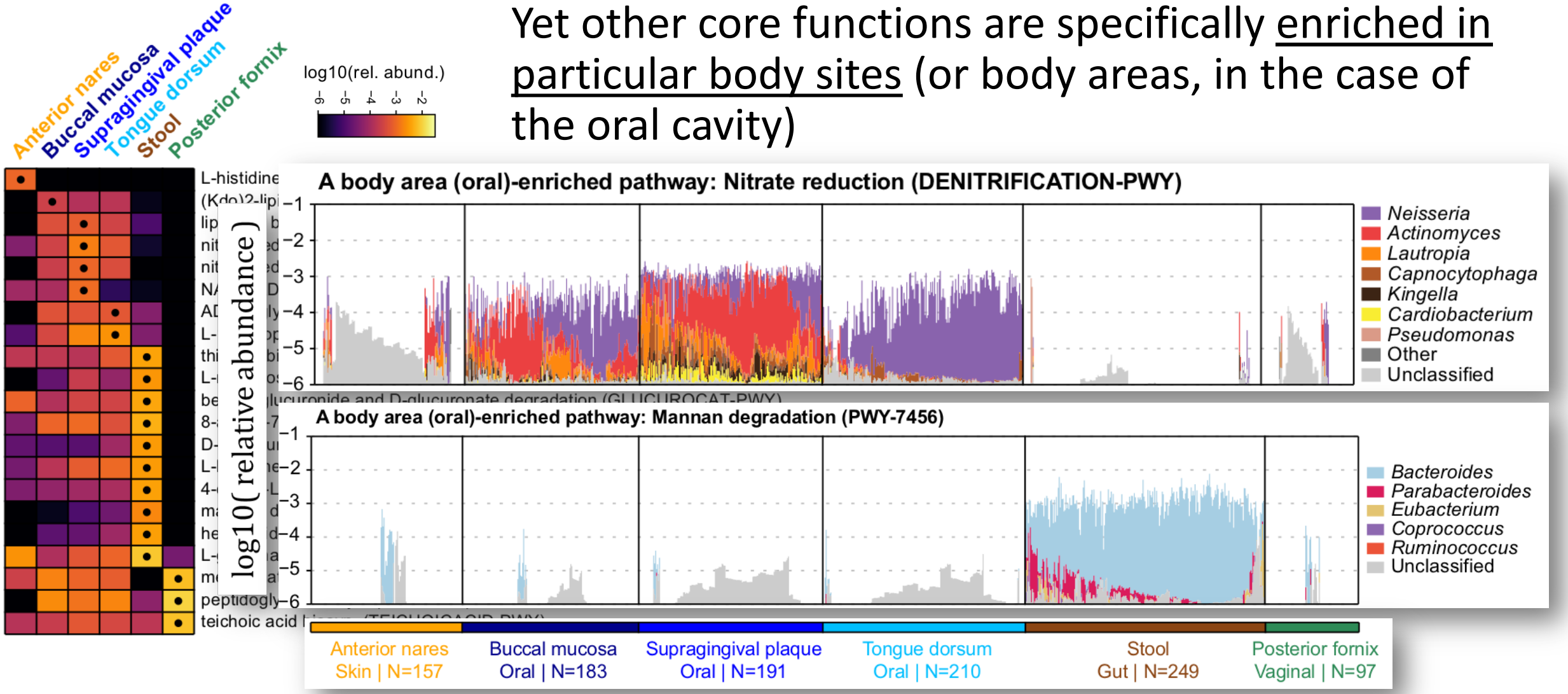


Yet other core functions are specifically <u>enriched in particular body sites</u> (or body areas, in the case of the oral cavity)

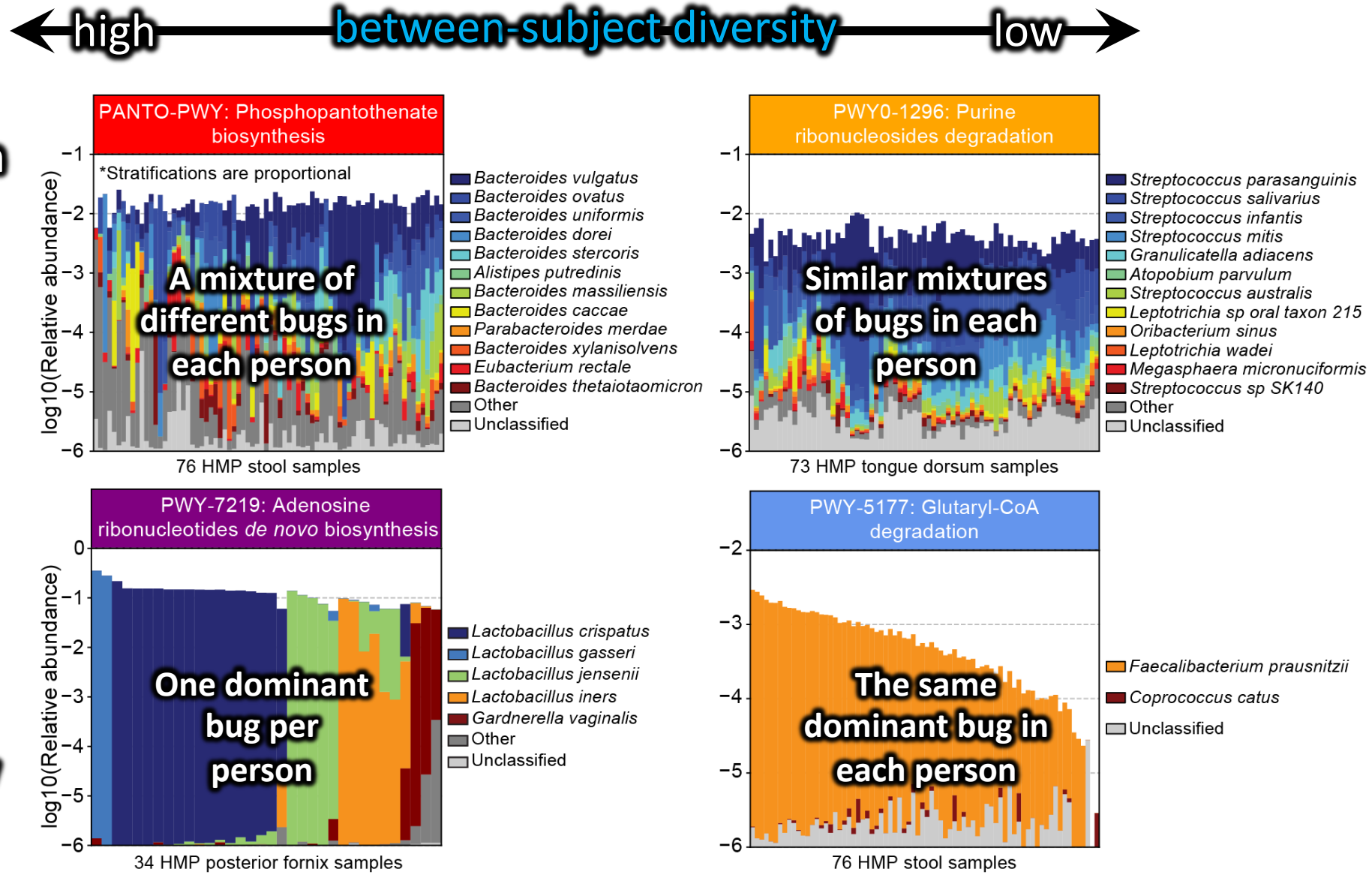
Core functions of the human microbiome

Lloyd-Price, Nature, 2017

Yet other core functions are specifically <u>enriched in particular body sites</u> (or body areas, in the case of the oral cavity)



Contributional diversity of core microbiome pathways



Apply community level diversity measures to species' contributes to individual functions

contributional diversity

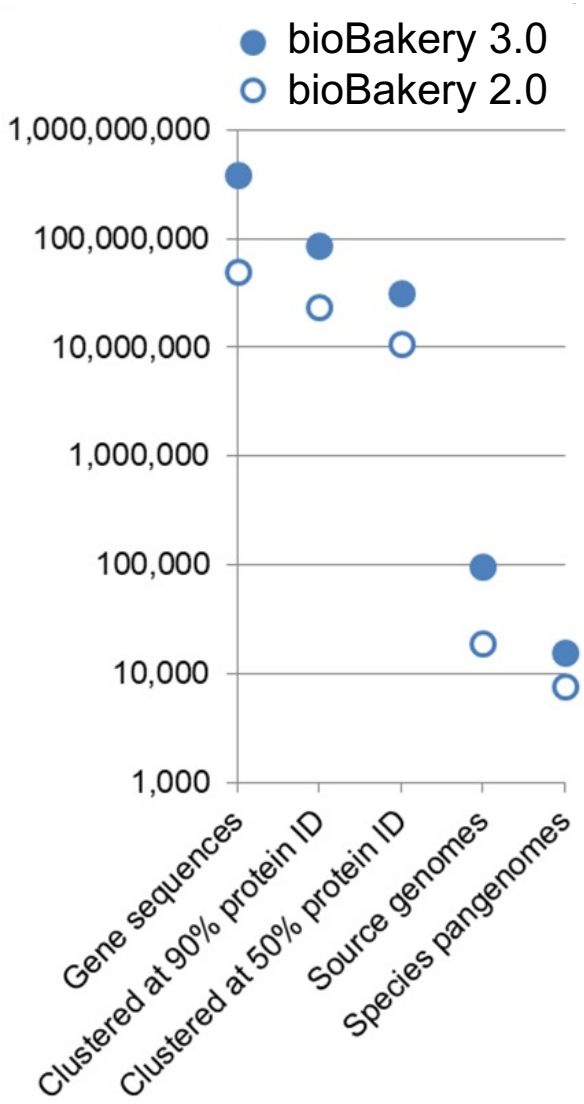
Within (alpha) and between (beta) sample flavors

http://huttenhower.sph.harvard.edu/humann

What's new in bioBakery 3.0?

(including HUMAnN 3.0)

bioBakery 3 builds bigger databases more easily



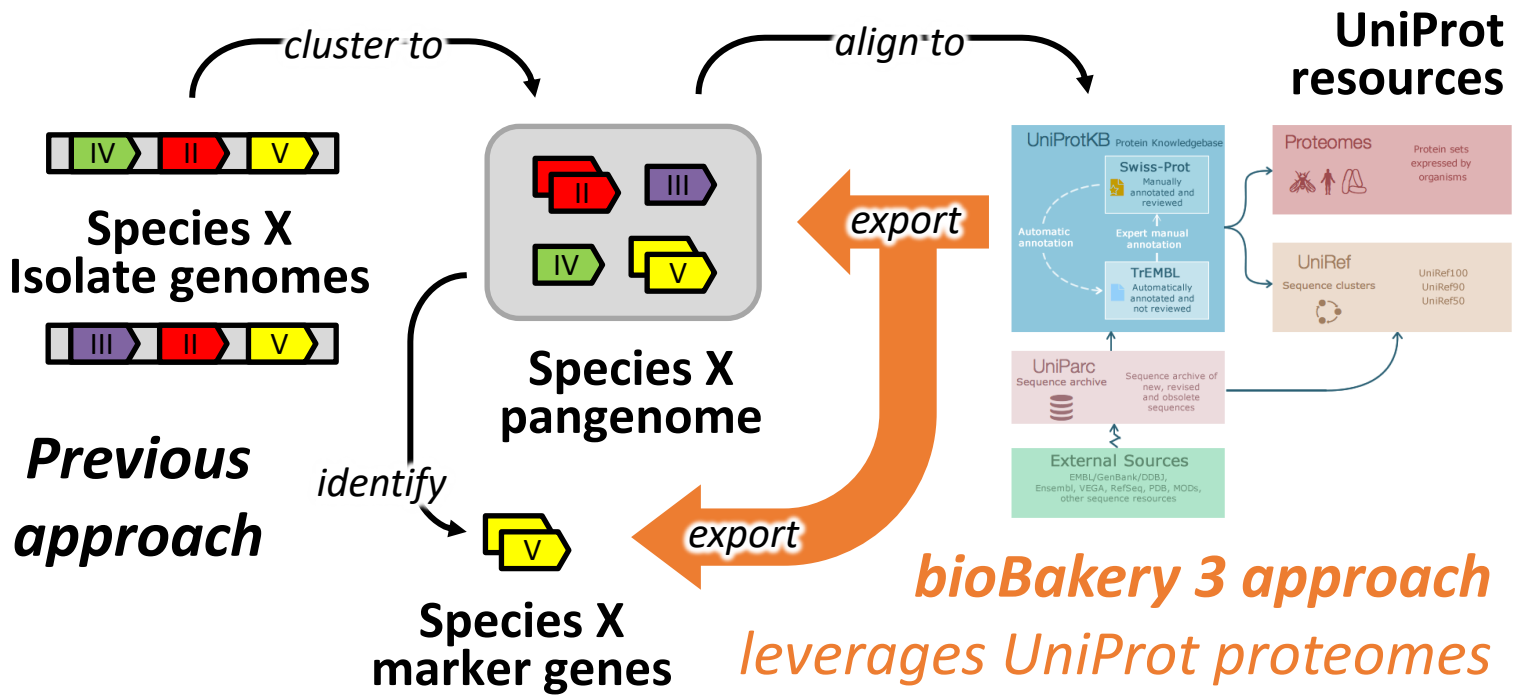
- 99K microbial genomes (**5x increase** vs. bioBakery 2)
- 16K microbial species (**3x increase**)
- 87M gene families (**3x increase**)
- 389M gene sequences (**8x increase**)

http://huttenhower.sph.harvard.edu/biobakery

bioBakery 3 builds bigger databases more easily

- 99K microbial genomes (**5x increase** vs. bioBakery 2)
- 16K microbial species (**3x increase**)
- 87M gene families (**3x increase**)
- 389M gene sequences (**8x increase**)

Facilitated by new methods for sequence database construction

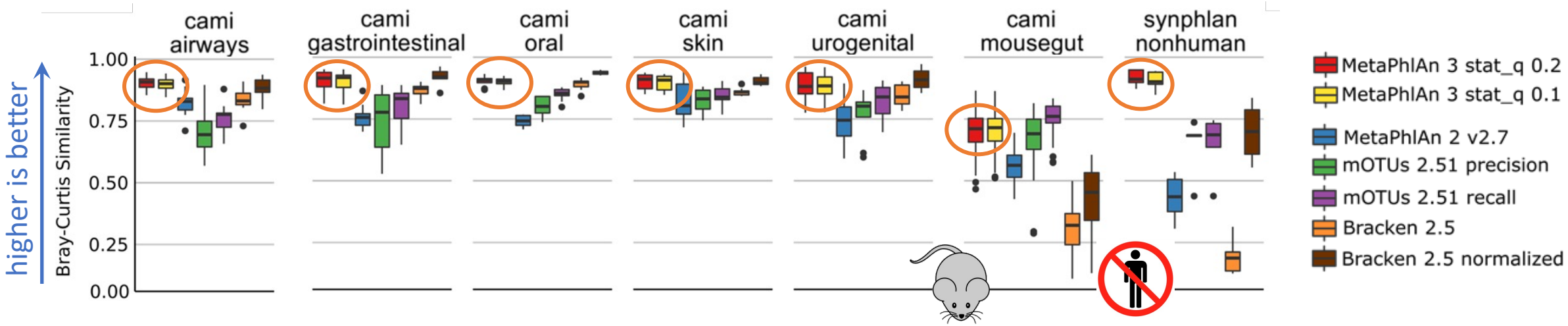


bioBakery 3 approach *leverages UniProt proteomes and UniRef families*

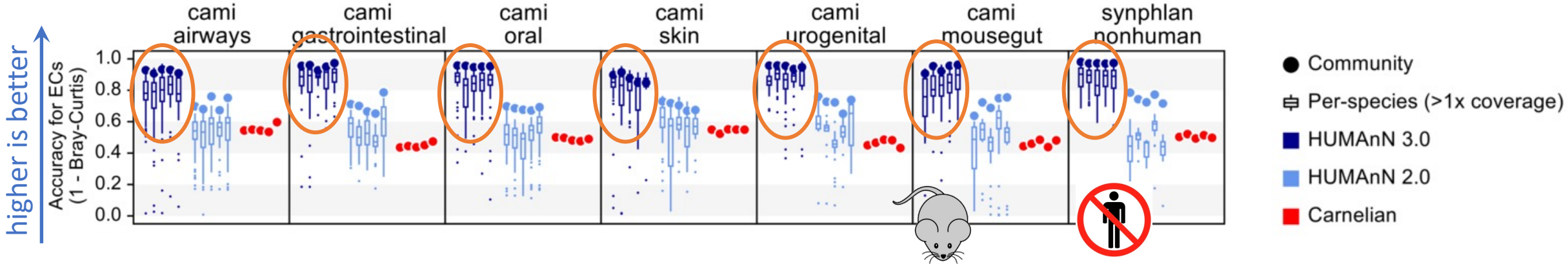
http://huttenhower.sph.harvard.edu/biobakery

bioBakery 3 accurately profiles synthetic communities

- MetaPhlAn 3 accurately profiles species from synthetic metagenomes (CAMI, etc.)



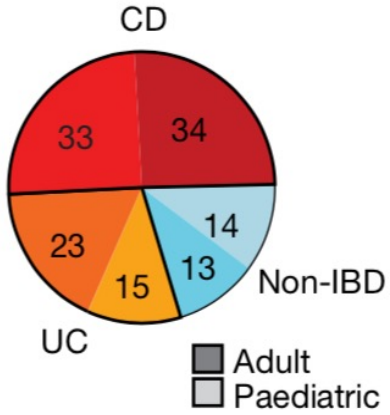
- HUMAnN 3 accurately profiles enzyme abundance (total and per-species)



Beghini et al. *eLife* 2021

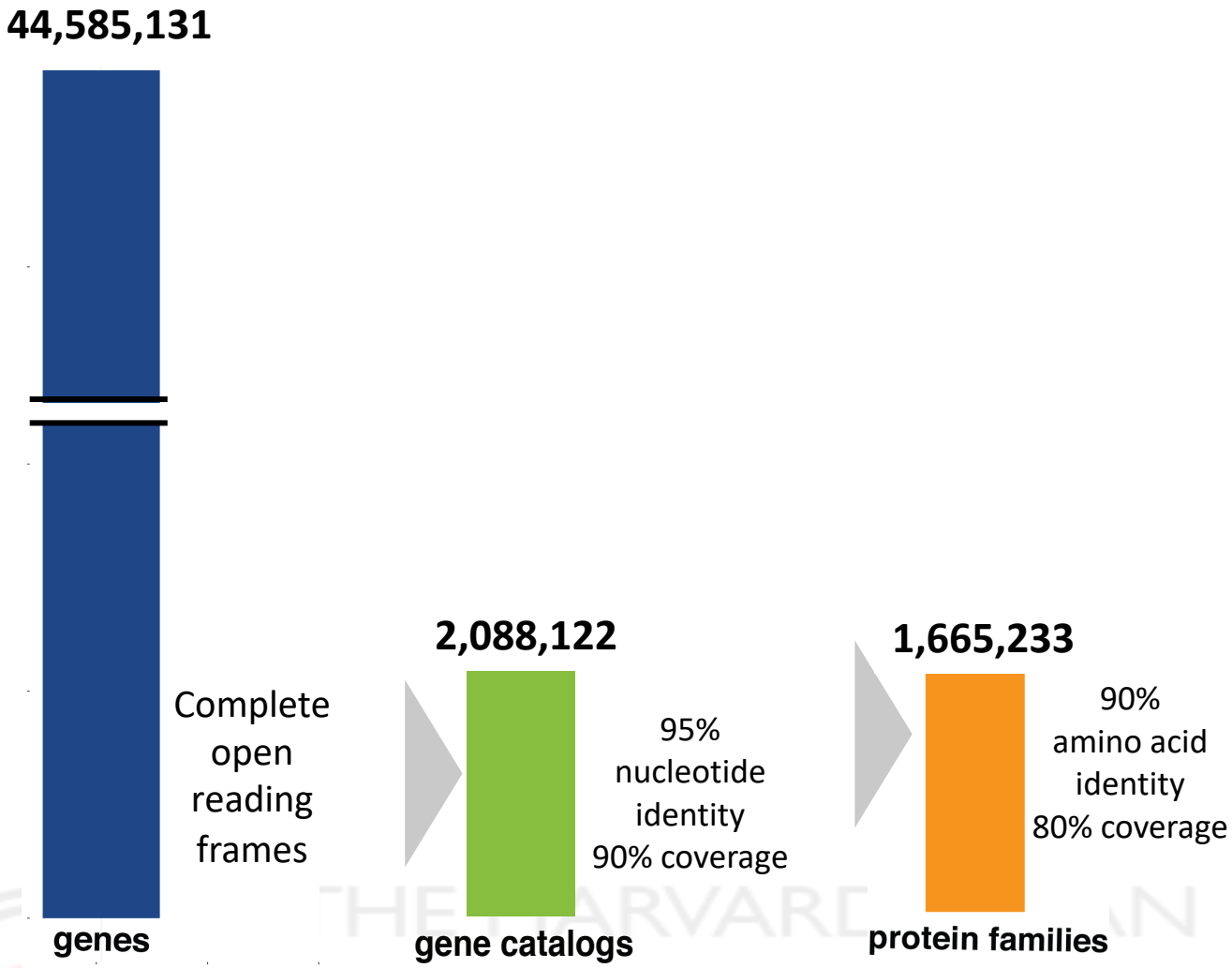
Prioritizing bioactive gene families with MetaWIBELE

Even in the human gut, uncharacterized proteins abound

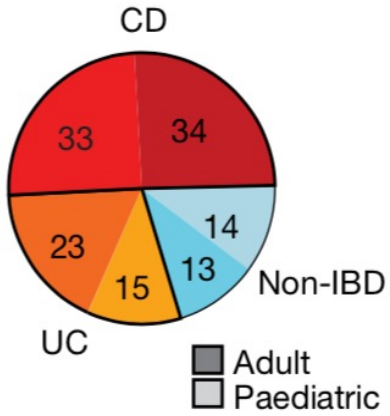


HMP2 Metagenomes (MGX)

- 132 participants' stool samples every two weeks for a year
- 1,595 total metagenomes

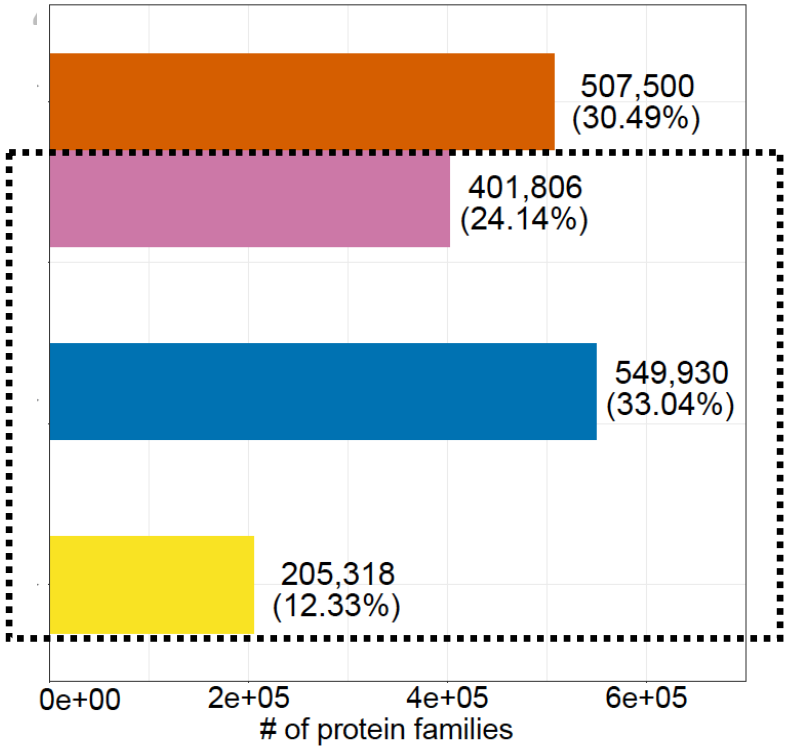


Even in the human gut, uncharacterized proteins abound



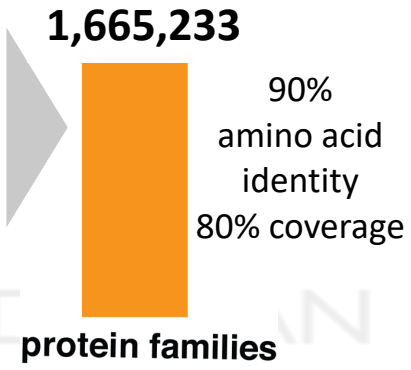
HMP2 Metagenomes (MGX)

- 132 participants' stool samples every two weeks for a year
- 1,595 total metagenomes



~70% of protein families uncharacterized

Strong homology to known characterized proteins
Strong homology to known uncharacterized proteins
Remote homology to known proteins
No homology to known proteins

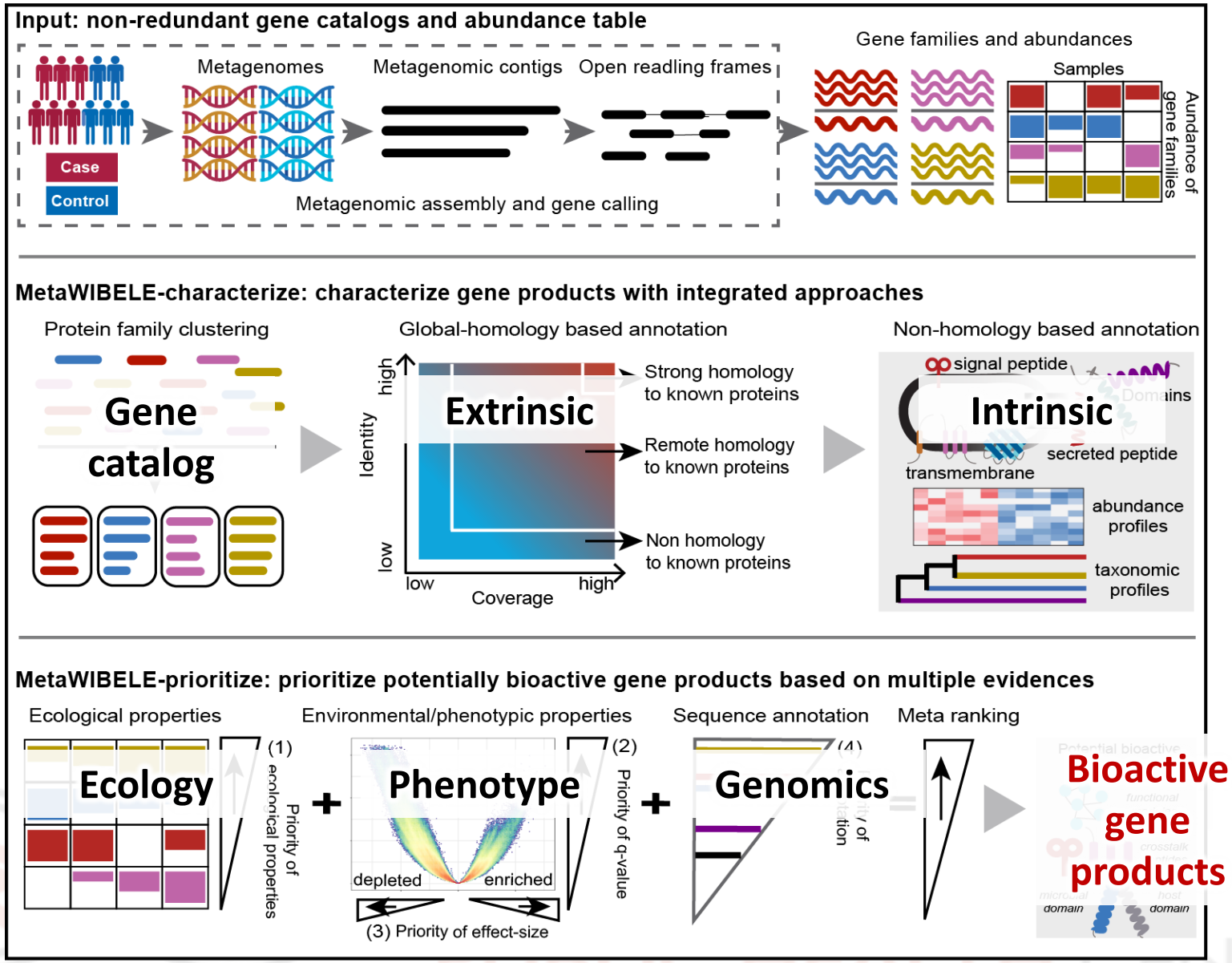


90%
amino acid
identity
80% coverage

Prioritizing and (initially) characterizing "important" microbial community proteins

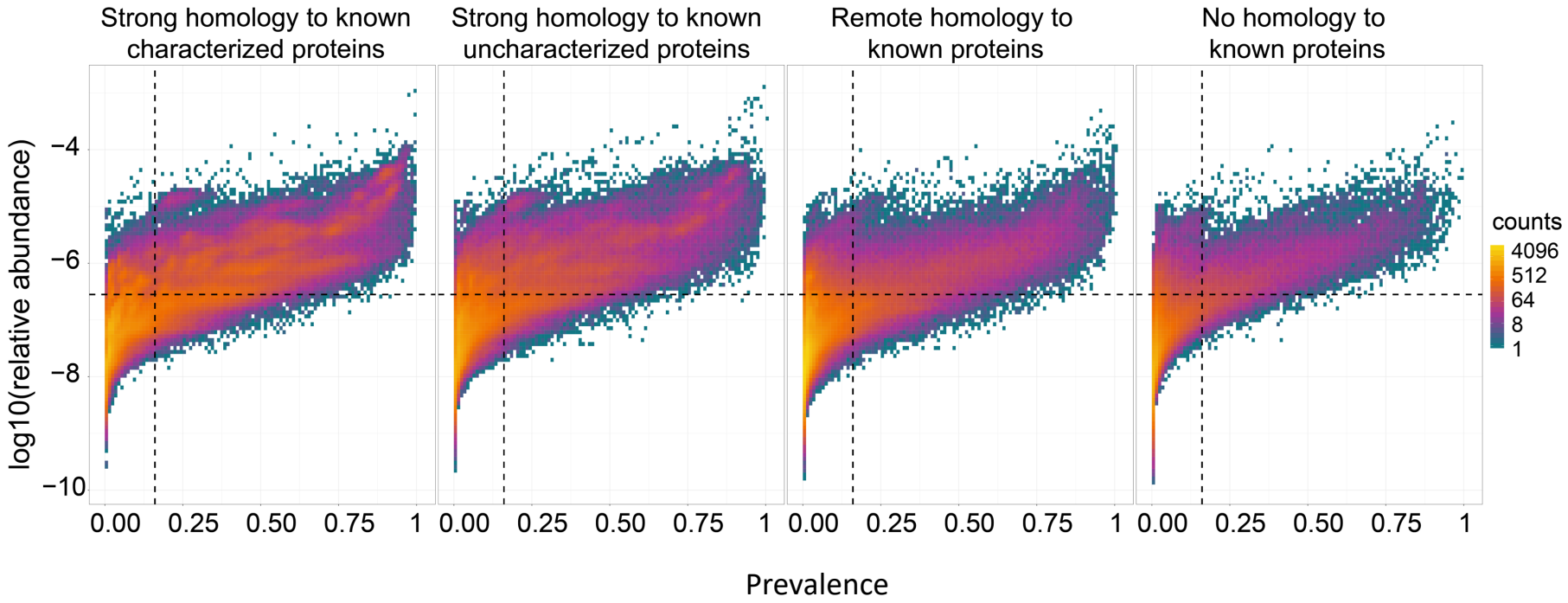
Yancong Zhang

MetaWIBELE: Workflow to Identify novel Bioactive Elements in the microbiome



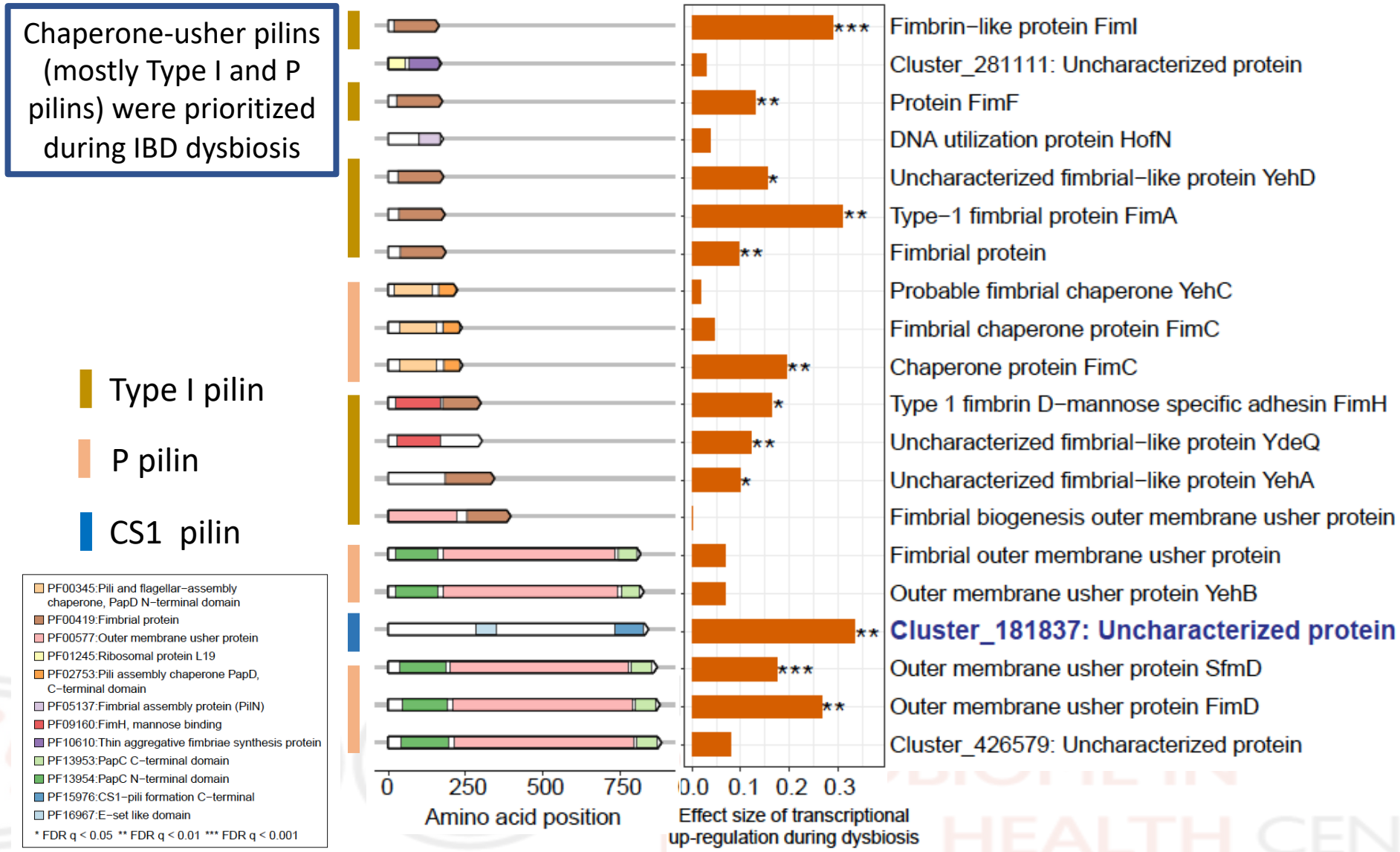
http://huttenhower.sph.harvard.edu/metawibele

Uncharacterized proteins are “typical”



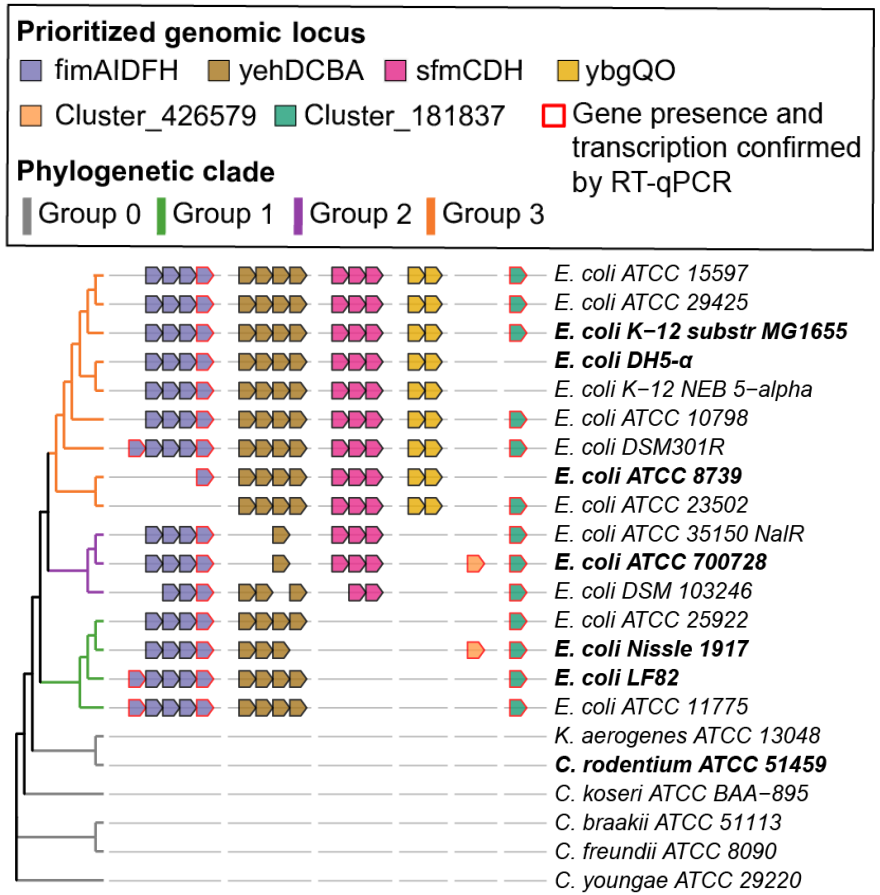
Bioactivity prioritizations in IBD are supported by external validation

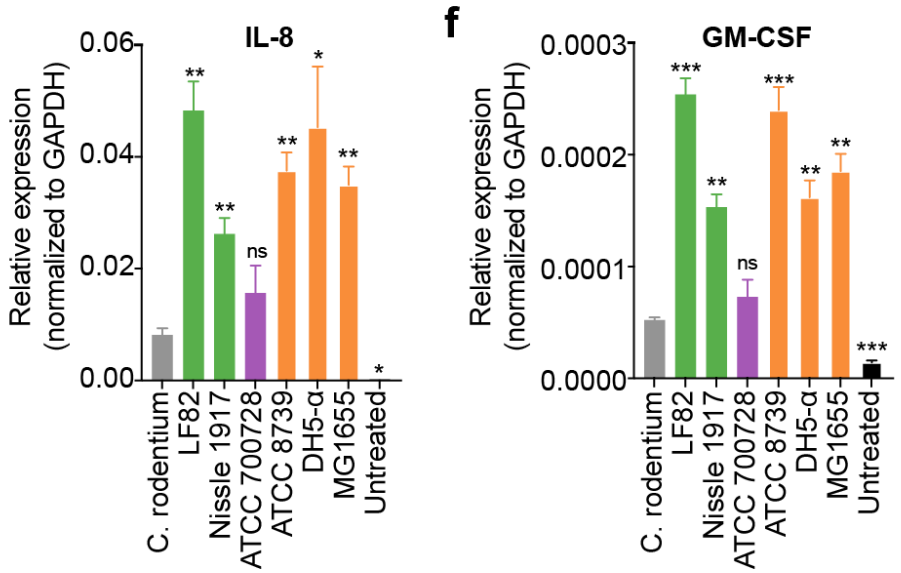
Chaperone-usher pilins (mostly Type I and P pilins) were prioritized during IBD dysbiosis



Bioactivity prioritizations in IBD are supported by external validation

Chaperone-usher pilins (mostly Type I and P pilins) were prioritized during IBD dysbiosis

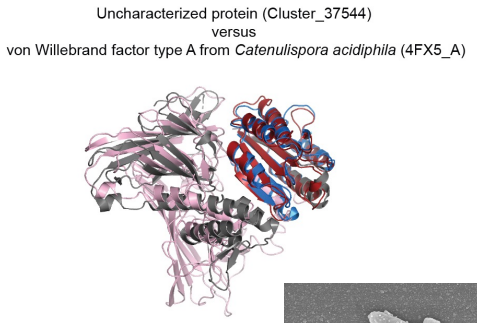


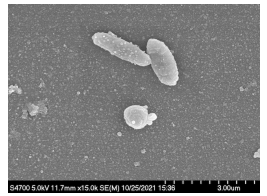


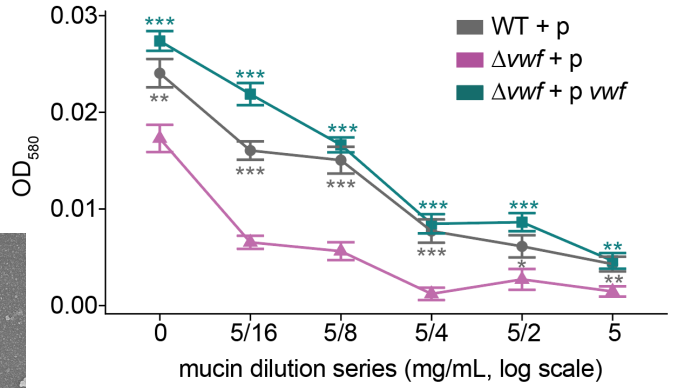
HCT-15 colonic epithelial cells
(MyD88 deficient, no flagellin response)

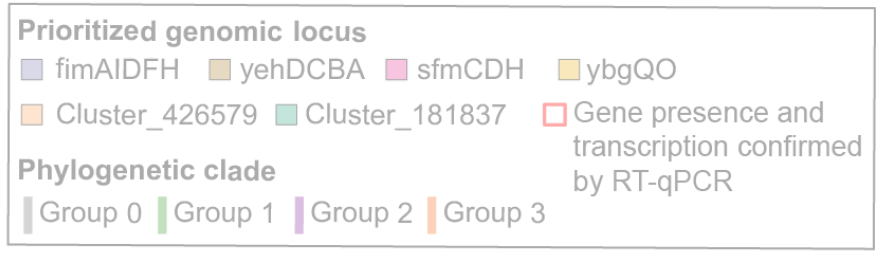
Bioactivity prioritizations in IBD are supported by external validation

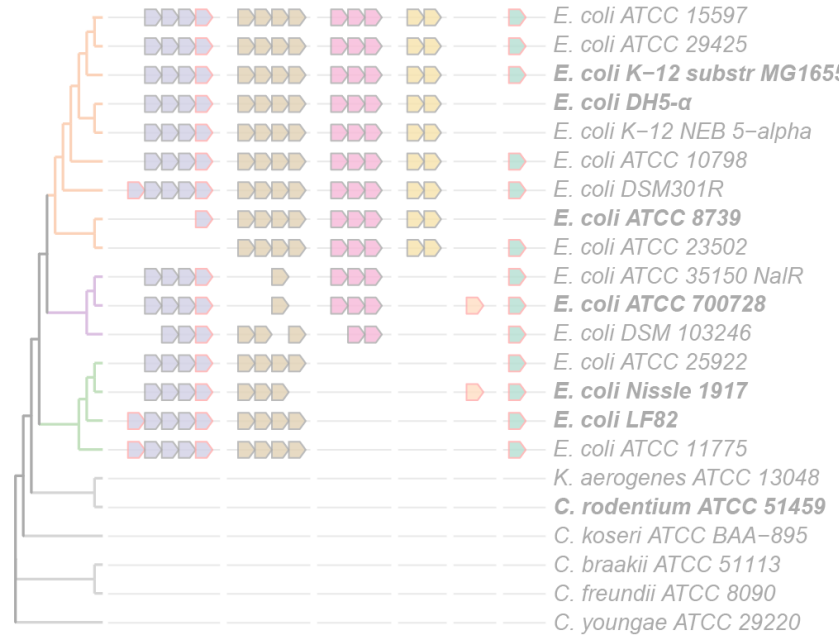
Chaperone-usher pilins (mostly Type I and P pilins) were prioritized during IBD dysbiosis

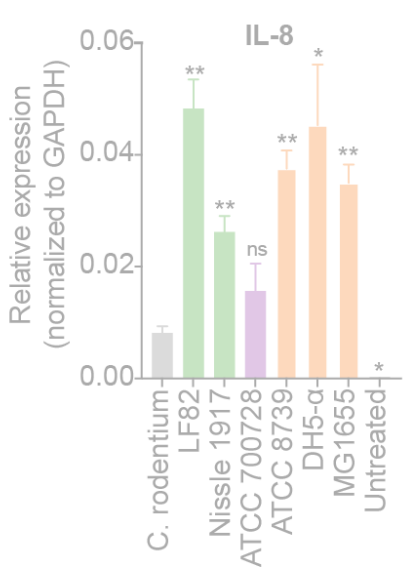


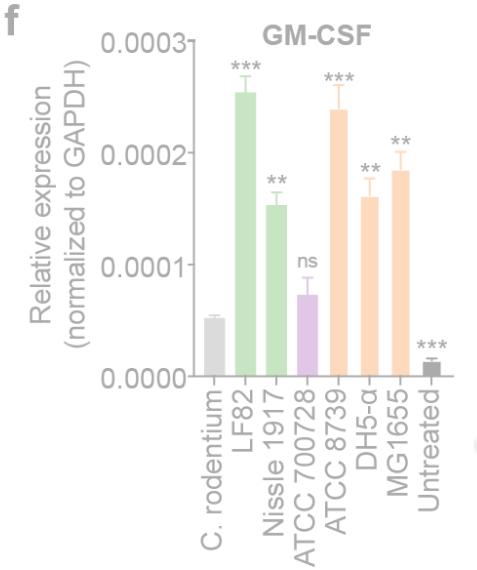












HCT-15 colonic epithelial cells
(MyD88 deficient, no flagellin response)

Summary

- Introduction to functional profiling
- Tiered search with HUMAnN
- Core functions of the human microbiome
- Contributional diversity
- What's new in bioBakery 3.0?
- Prioritizing bioactive gene families with MetaWIBELE

Microbial community functional profiling

(for real this time)

Summary

- Metatranscriptomics (MTX)
  - Functional profiling of MTX data
  - Relationships between functional potential and activity
  - Detecting under- and over-expression
  - Statistical models for MTX analysis
- Metabolomics (MBX)
  - Sequencing is easy… everything else is hard
  - Metabolomics approaches
  - Identifying compounds
  - Bioactivity prediction with MACARRoN

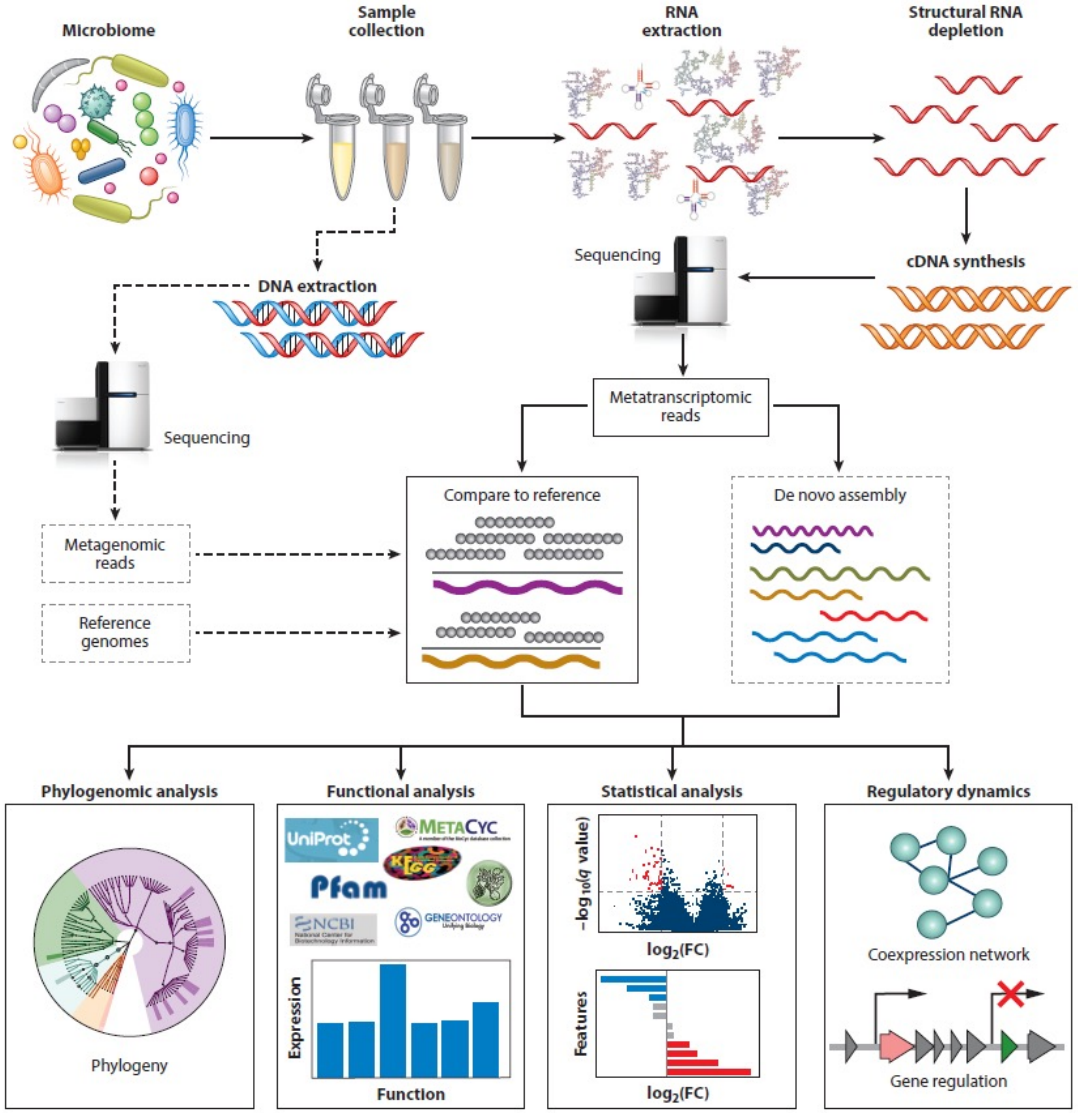
Metatranscriptomics (MTX) overview

ANNUAL REVIEWS

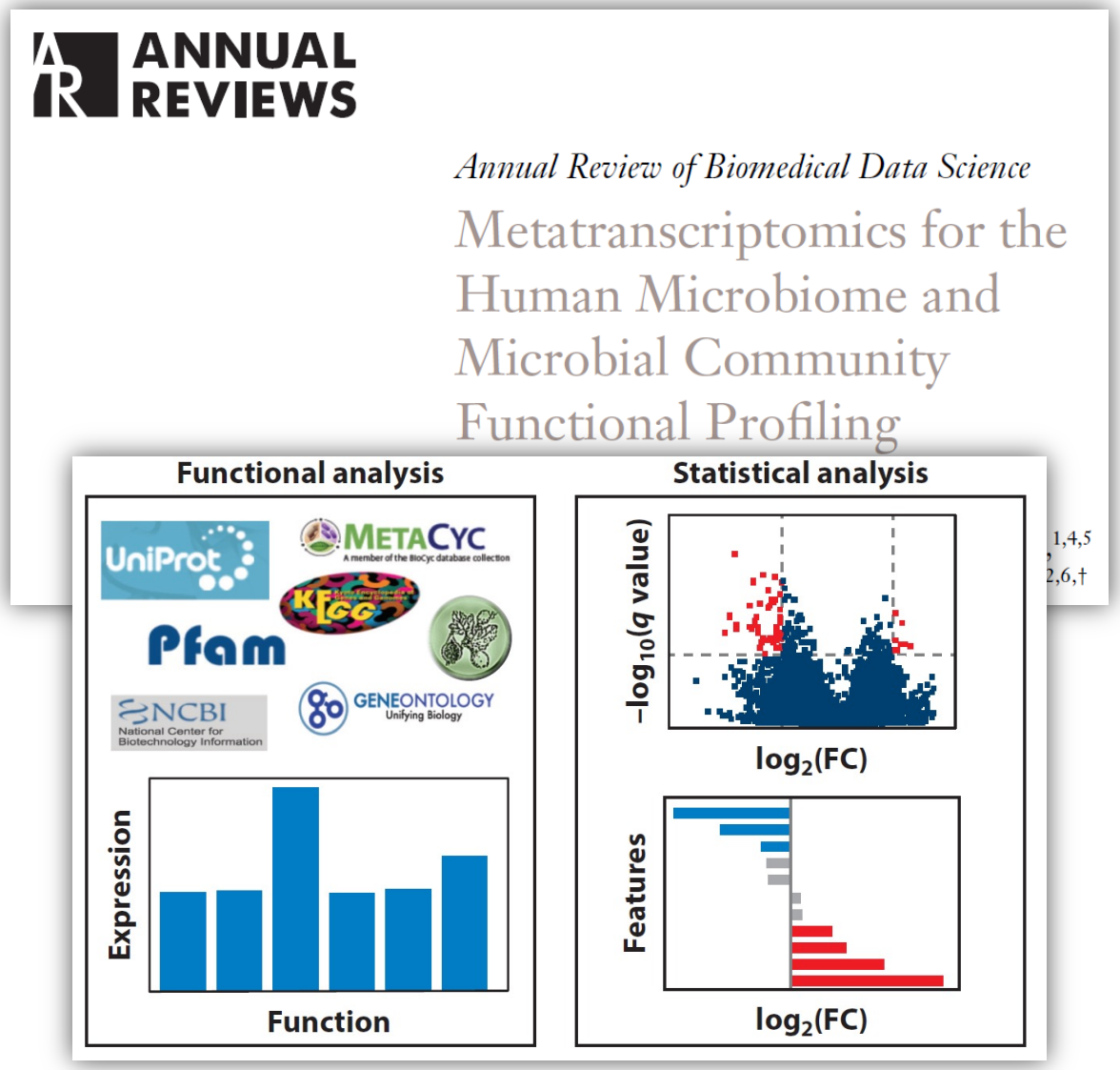
Annual Review of Biomedical Data Science

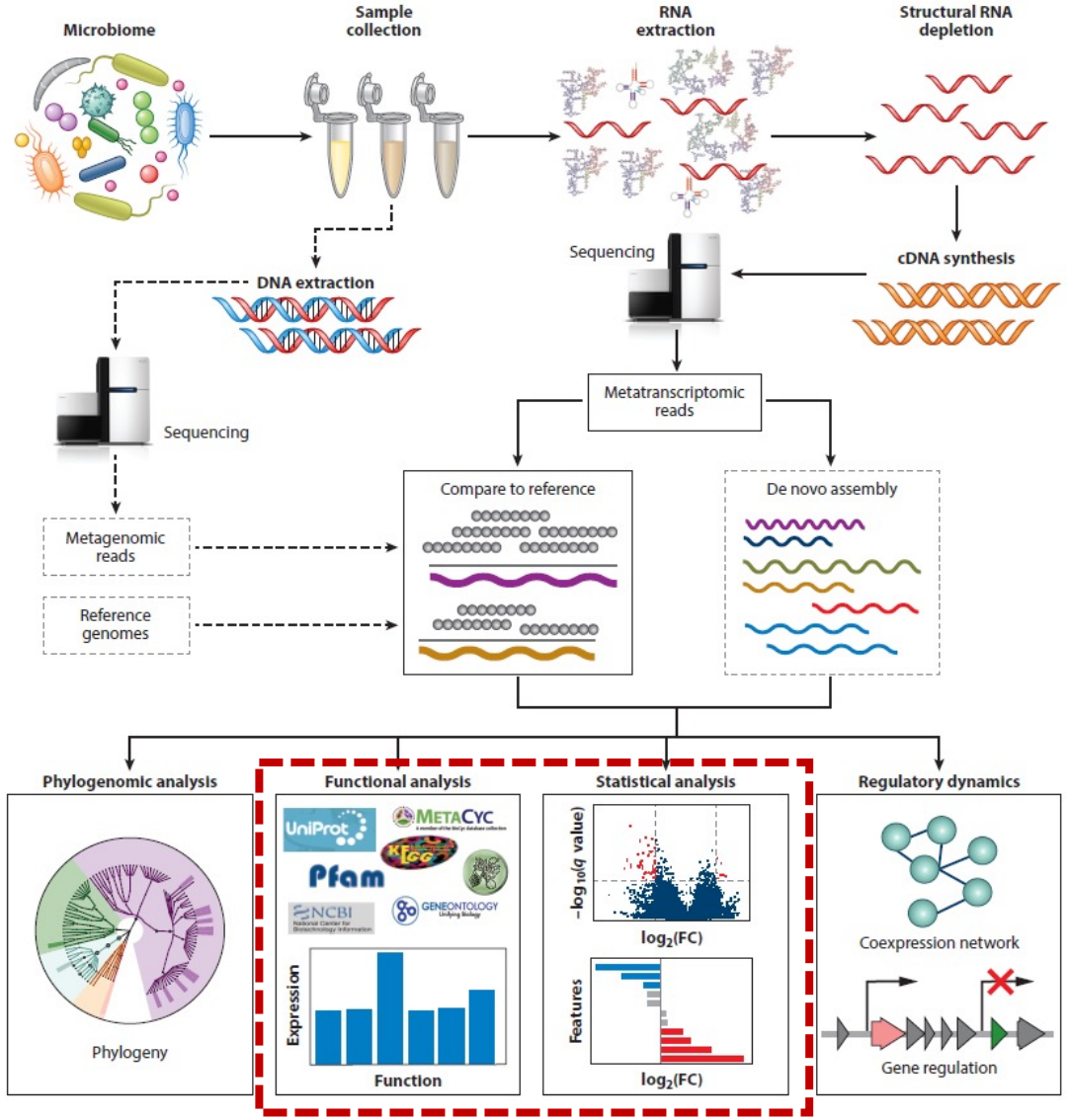
Metatranscriptomics for the Human Microbiome and Microbial Community Functional Profiling

Yancong Zhang,[1,2,*] Kelsey N. Thompson,[1,2,*] Tobyn Branck,[1,2,3] Yan Yan,[1,2] Long H. Nguyen,[1,4,5] Eric A. Franzosa,[1,2,†] and Curtis Huttenhower[1,2,6,†]

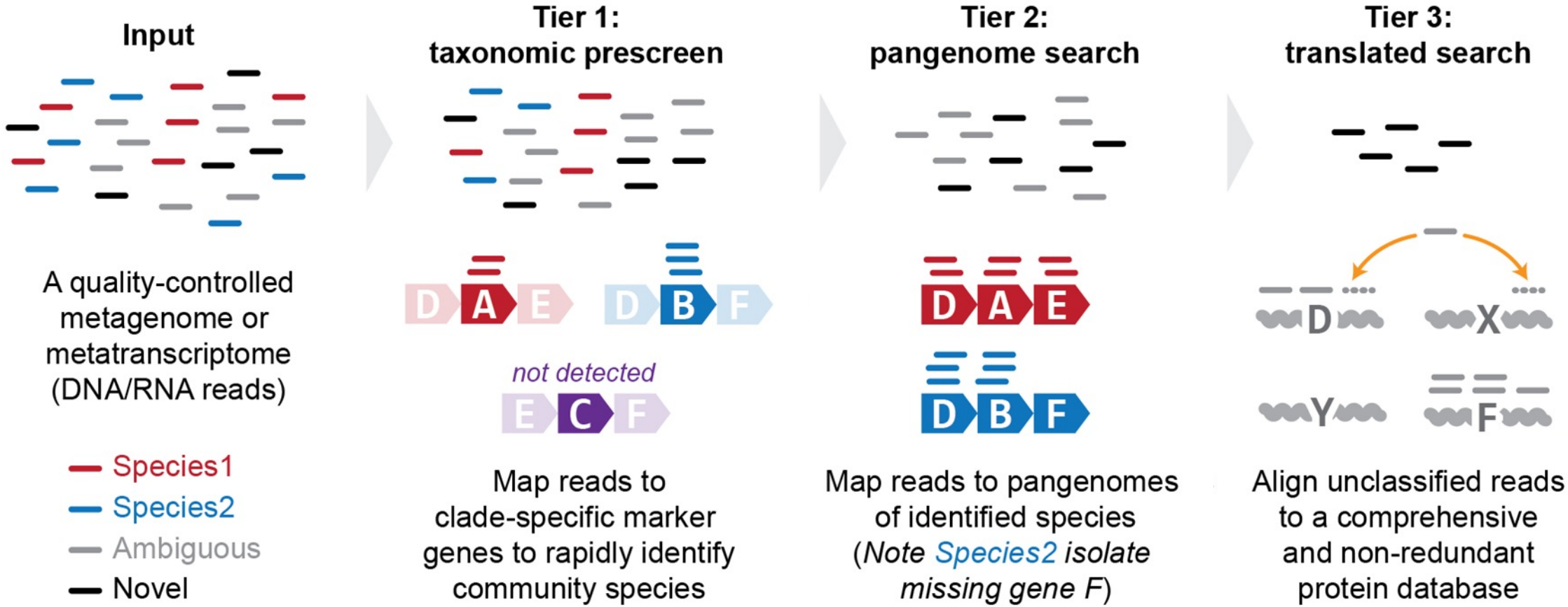


Metatranscriptomics (MTX) overview





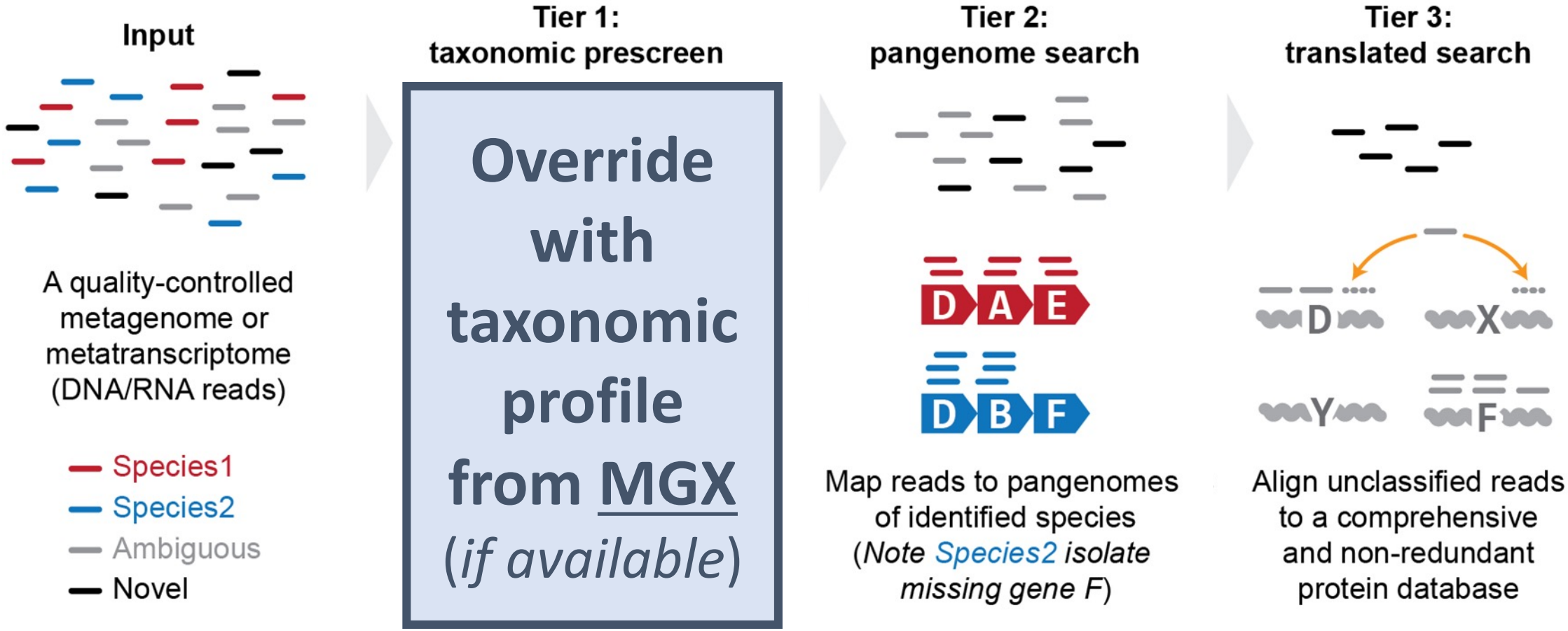
HUMAnN works on metatranscriptomes!



See **KneadData** for meta'omic QC

MetaPhlAn **HUMAnN**

HUMAnN works on metatranscriptomes!

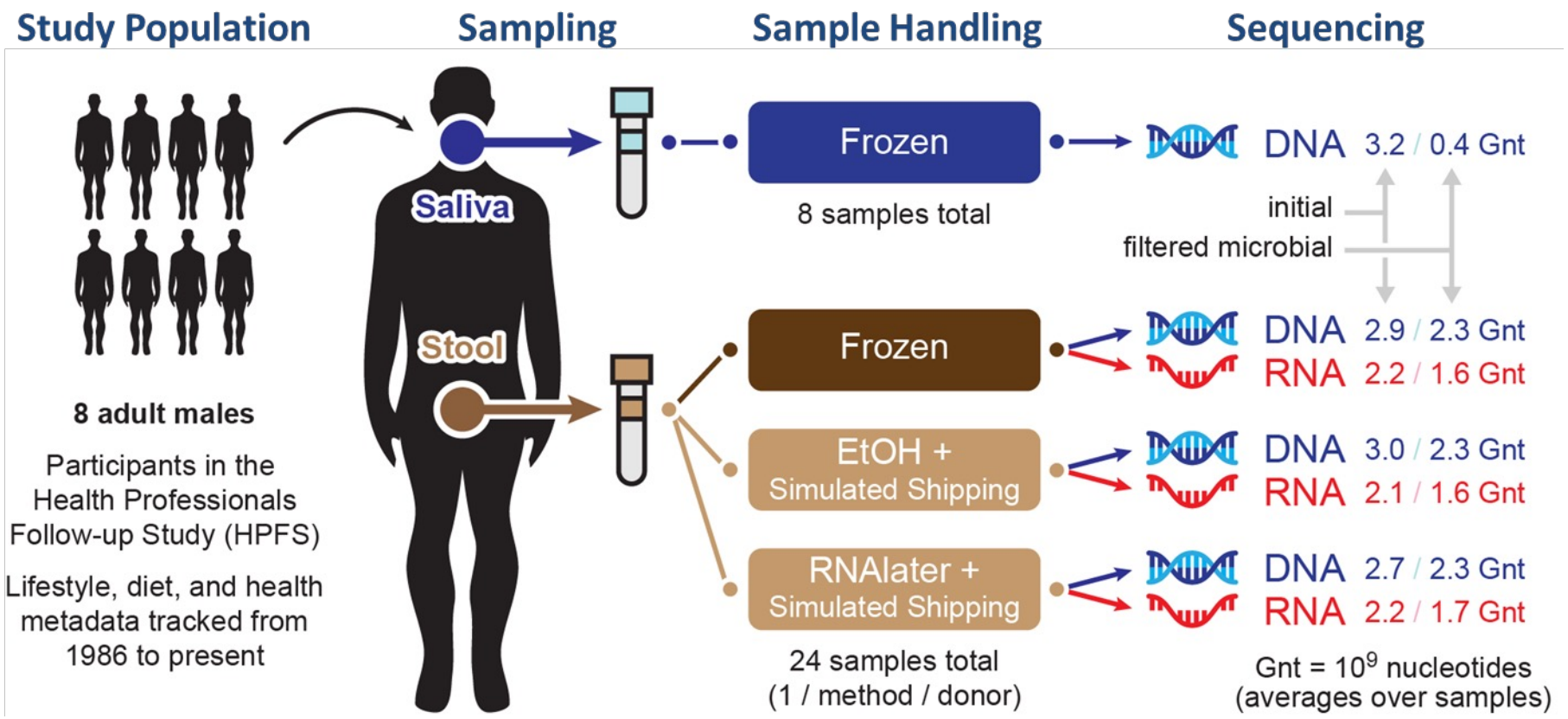


See **KneadData** for meta'omic QC

MetaPhlAn **HUMAnN**

Piloting metatranscriptomics

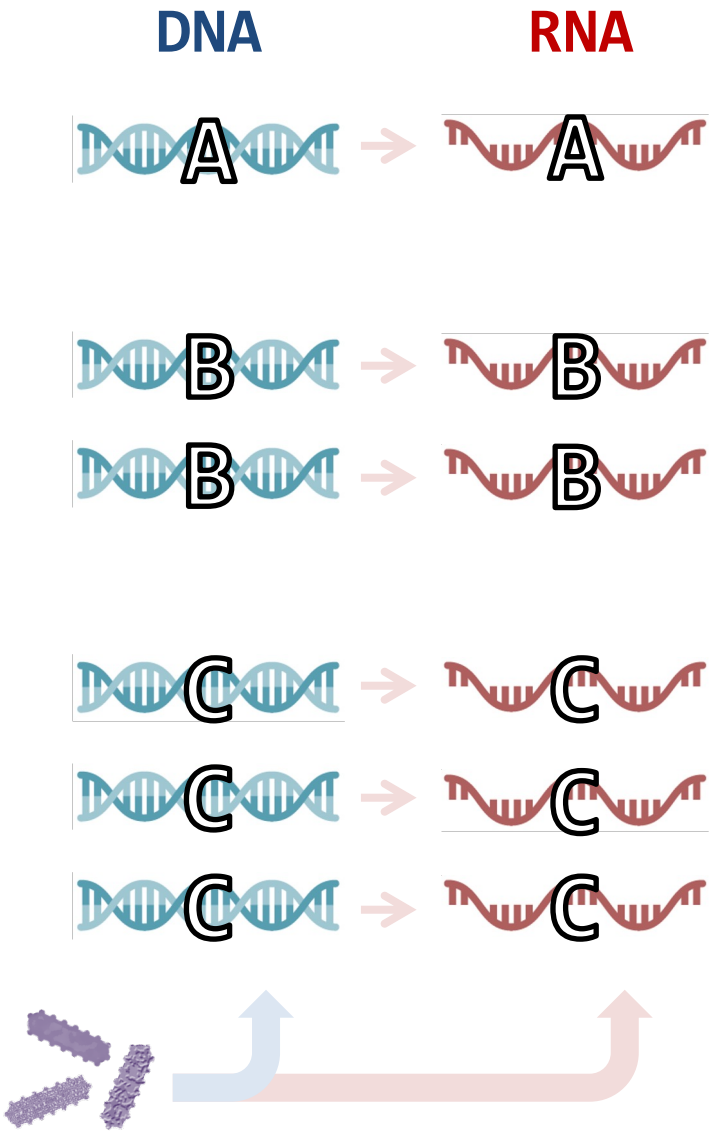
The HPFS pilot: self-collected MGX/MTX samples



8 self-collected stool samples were mock-shipped with 3 different preservation methods
Then they were subjected to paired MGX and MTX sequencing

Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Functional potential (MGX/DNA) vs. activity (MTX/RNA)



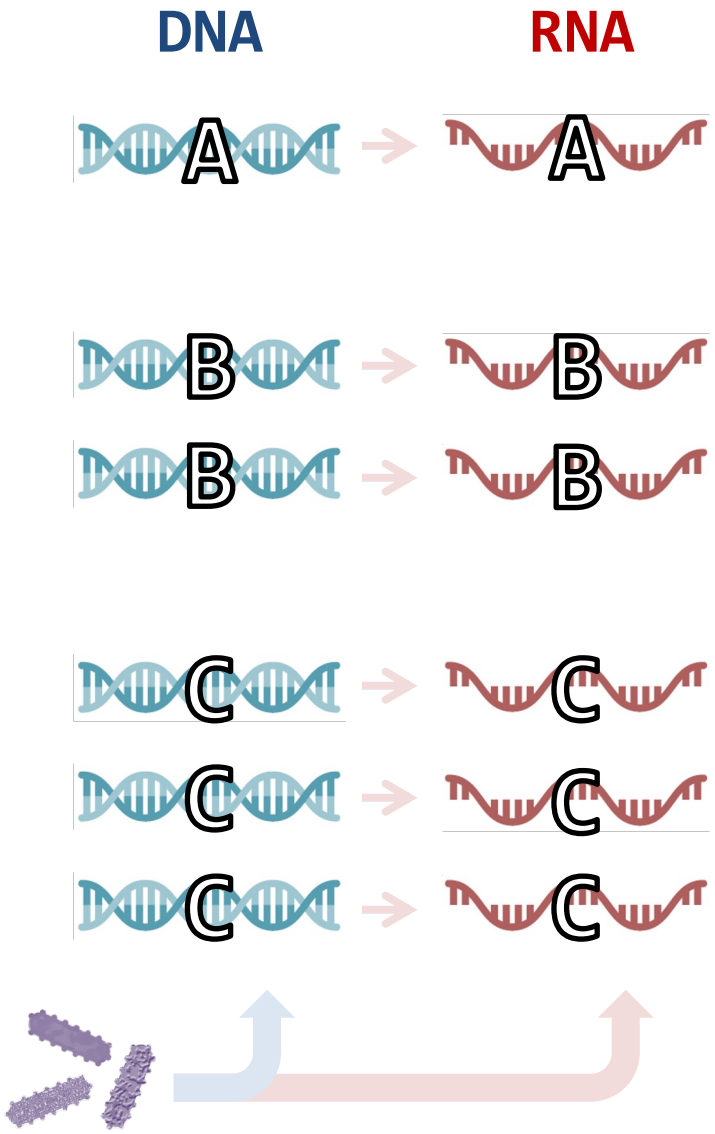
DNA data (metagenomics)
describes functional ***potential***

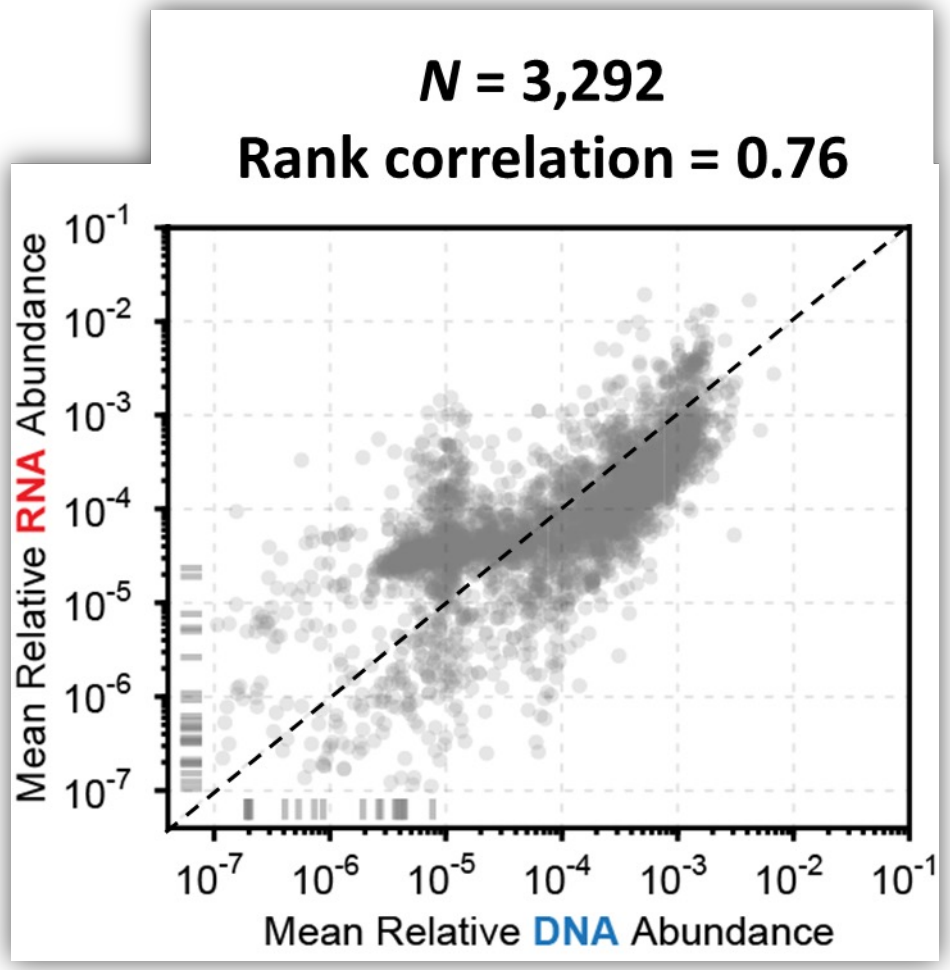
RNA data (metatranscriptomics)
describes functional ***activity***

Does ***potential*** = ***activity***
in the human gut?

Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

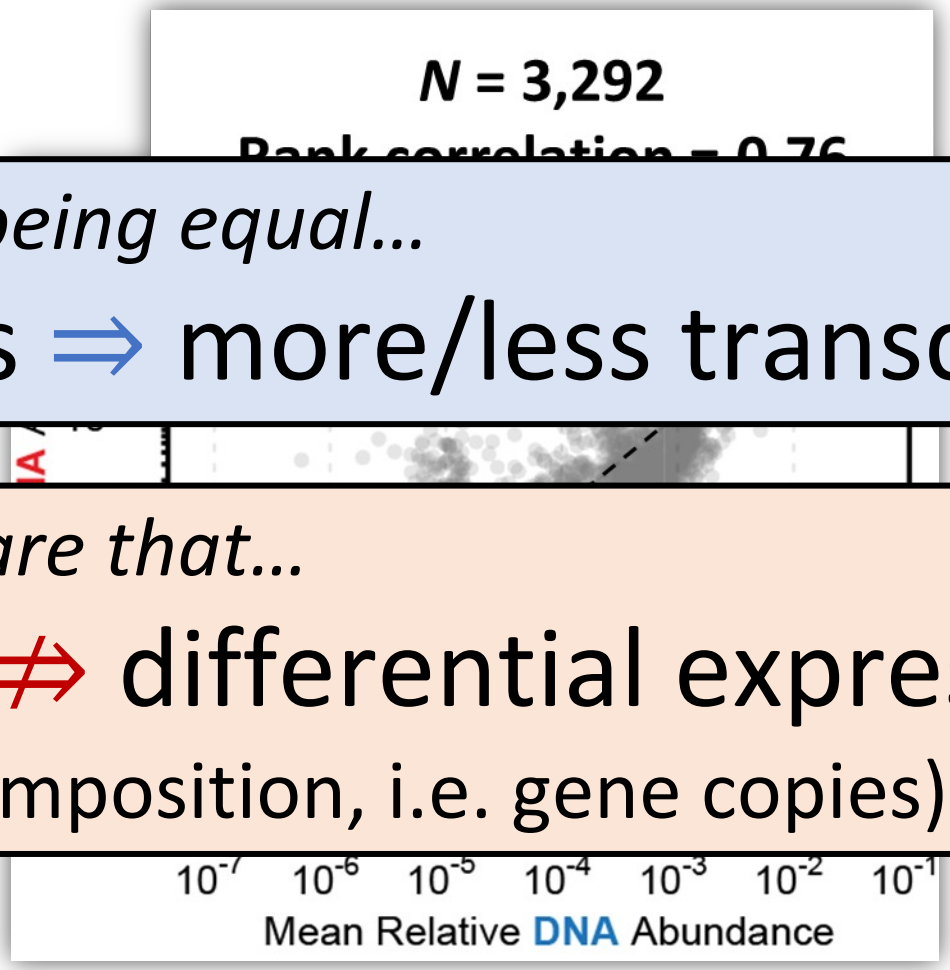
Potential and activity are strongly correlated!





Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Potential and activity are strongly correlated!



all else being equal…

more/less gene copies ⇒ more/less transcripts

beware that…

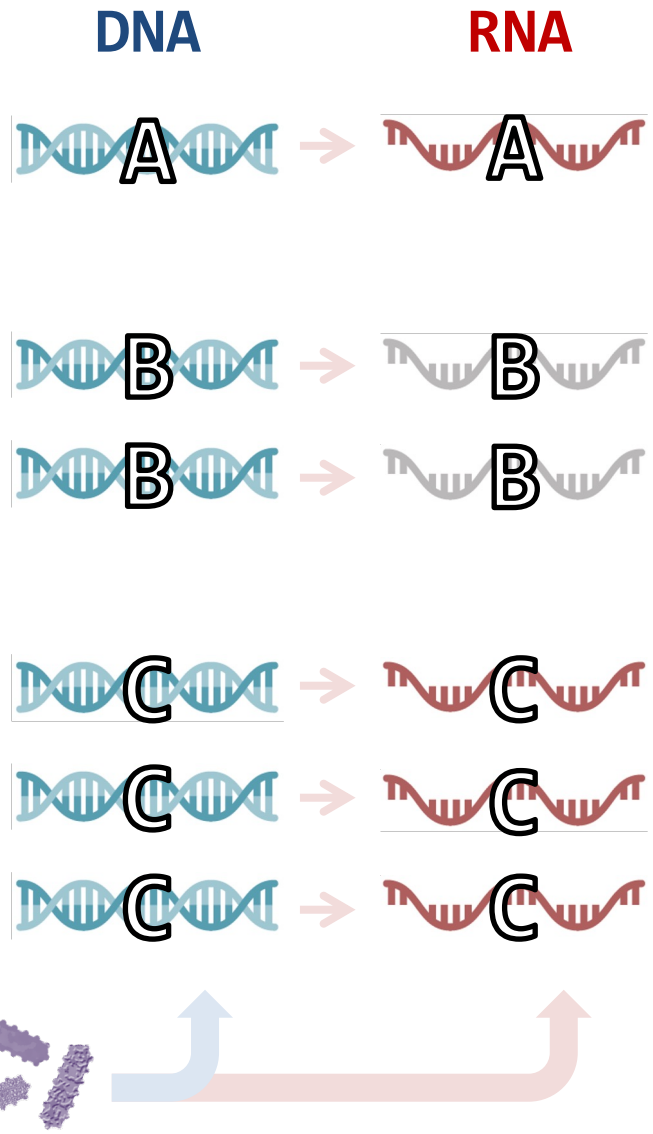
more/less transcripts ⇏ differential expression

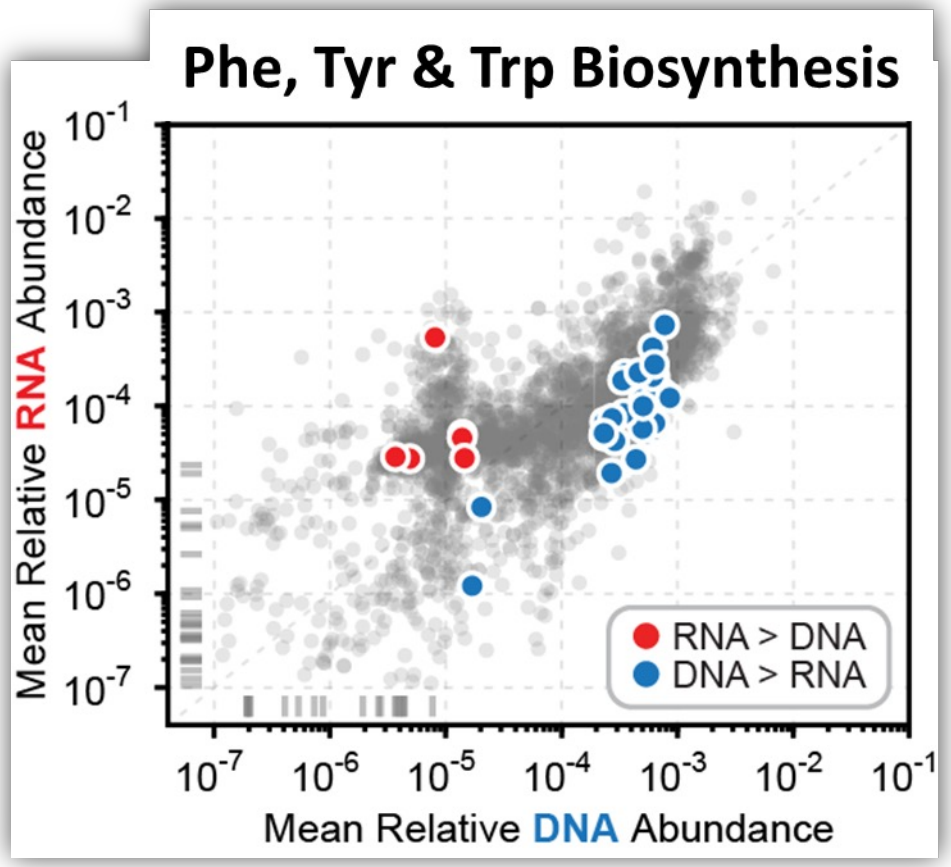
(could be changes in composition, i.e. gene copies)

Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Some functions are “under-abundant” in MTX

Low RNA abundance relative to DNA abundance

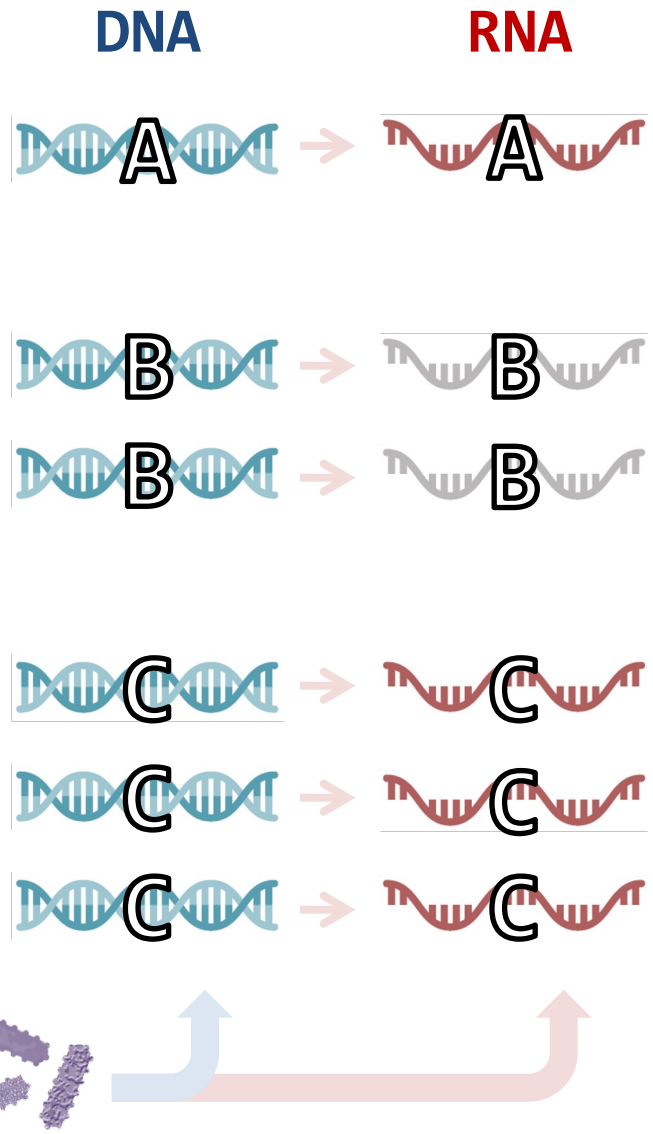


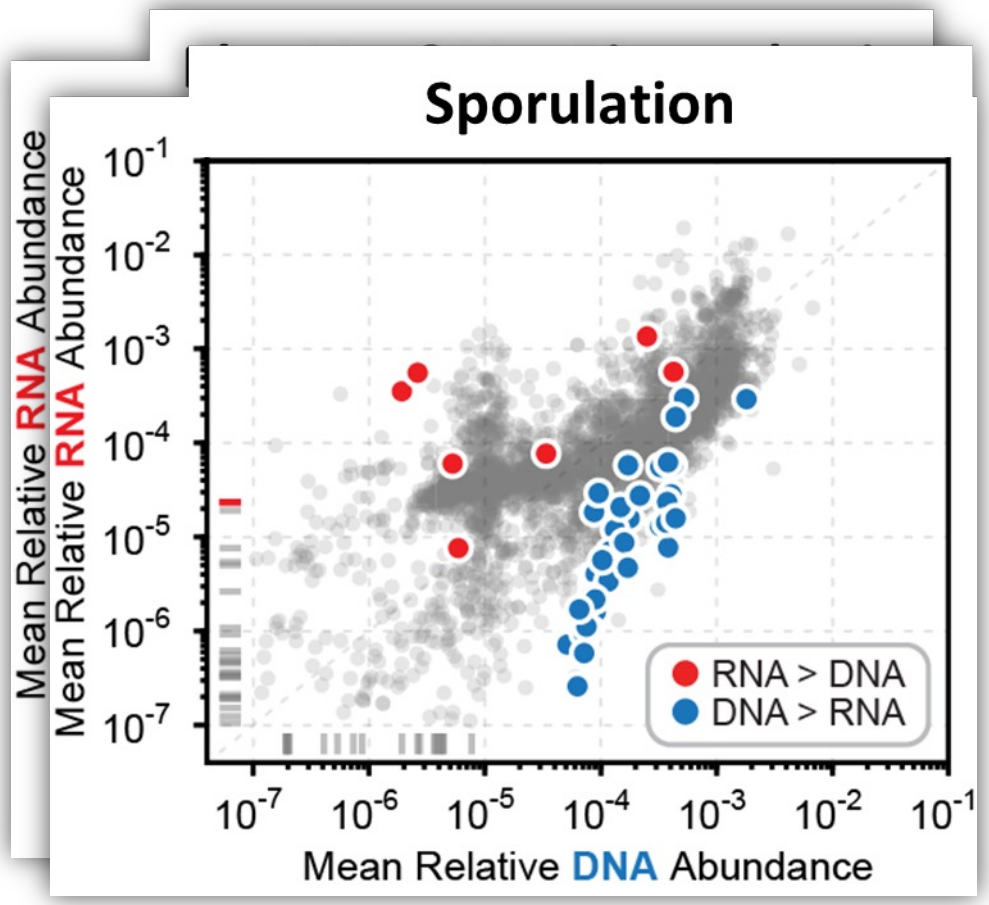


Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Some functions are “under-abundant” in MTX

Low RNA abundance relative to DNA abundance

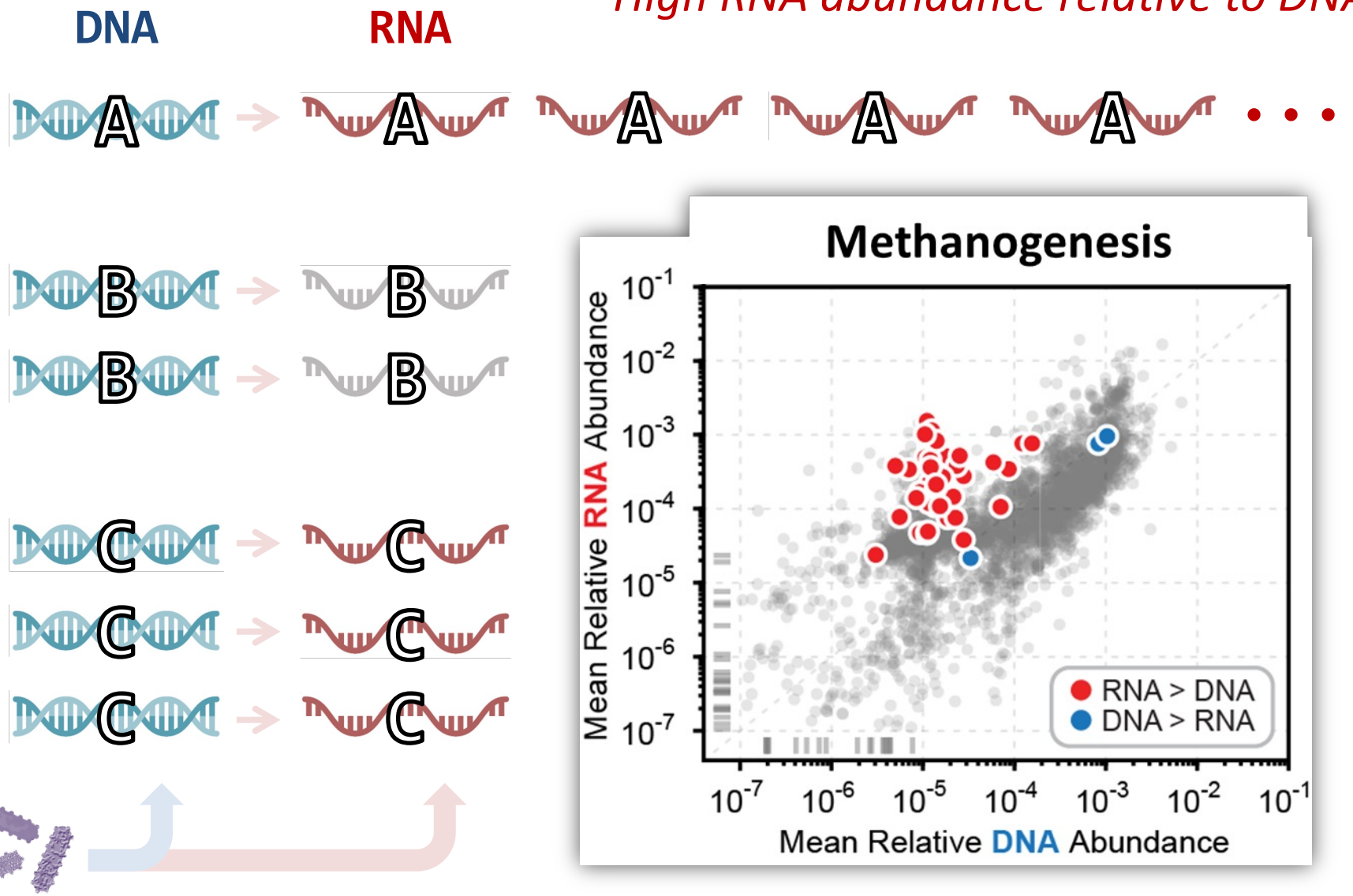




Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Other functions are “over-abundant” in MTX

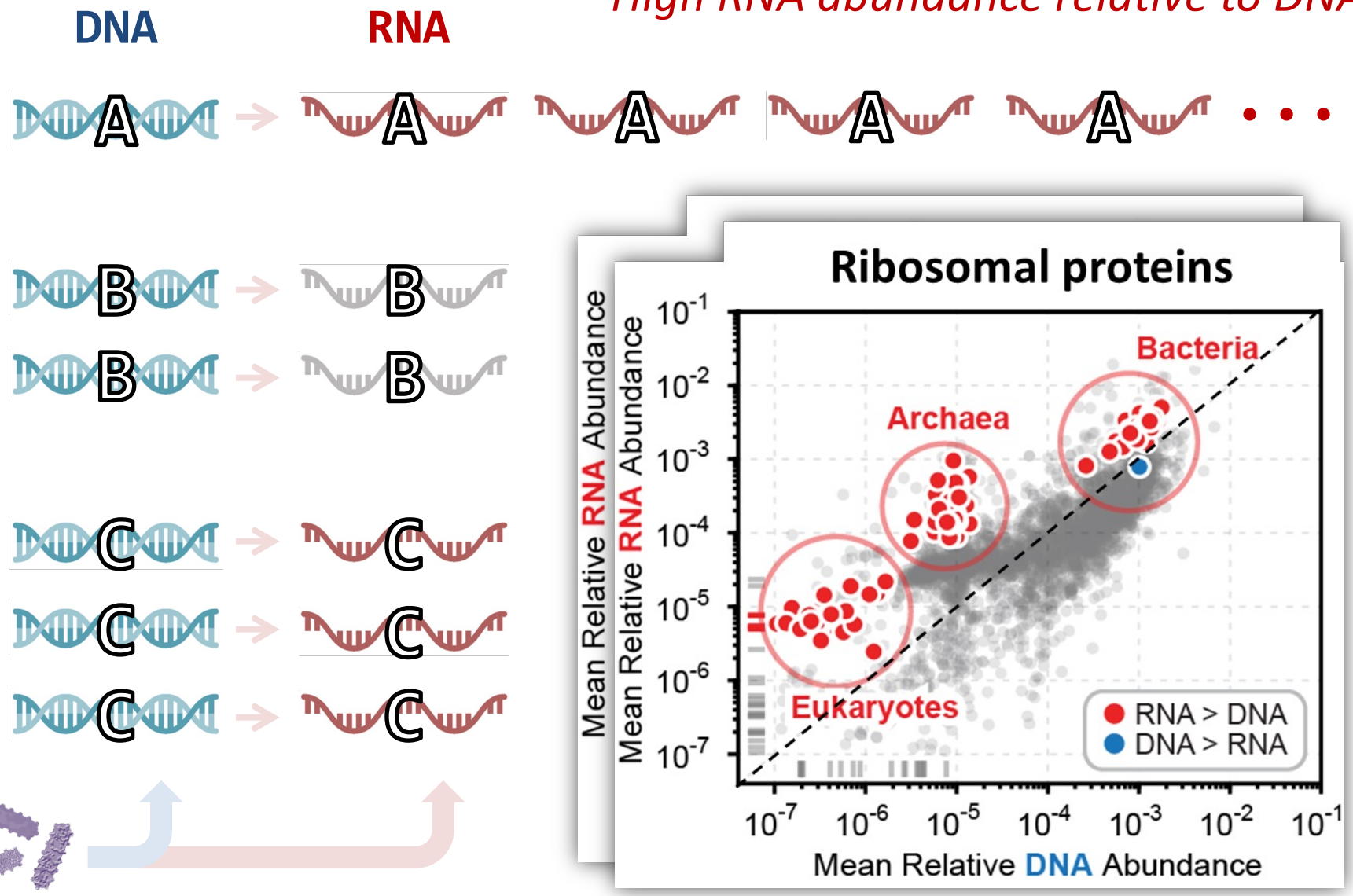
High RNA abundance relative to DNA abundance



Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Other functions are “over-abundant” in MTX

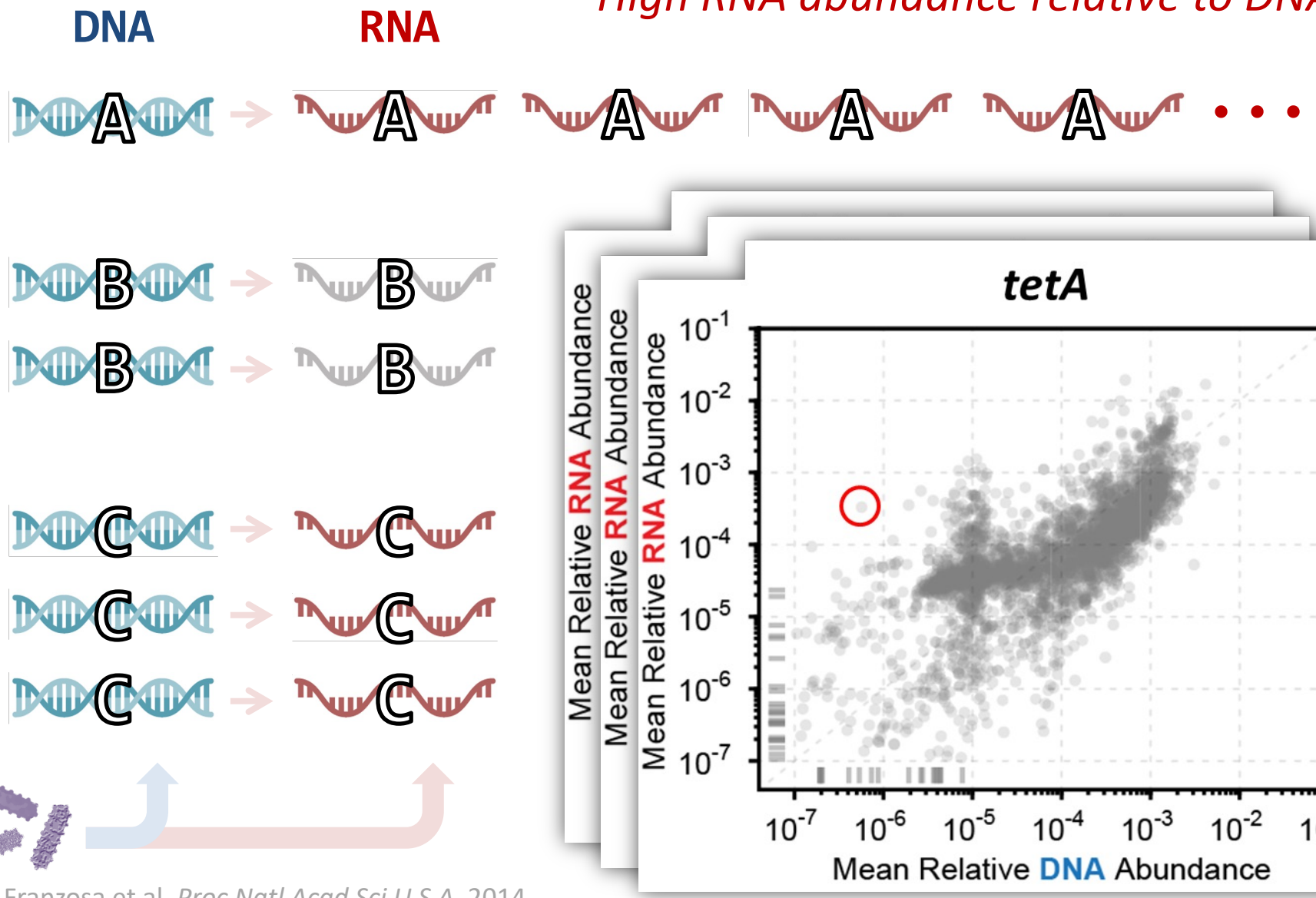
High RNA abundance relative to DNA abundance



Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

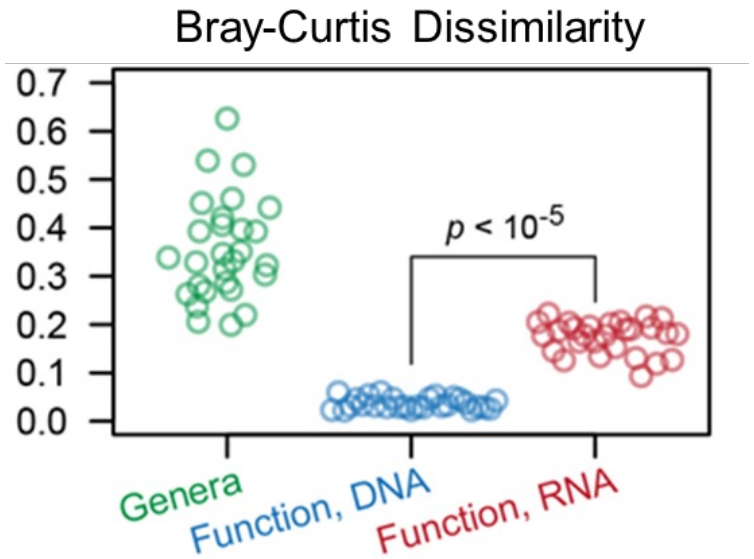
Other functions are “over-abundant” in MTX

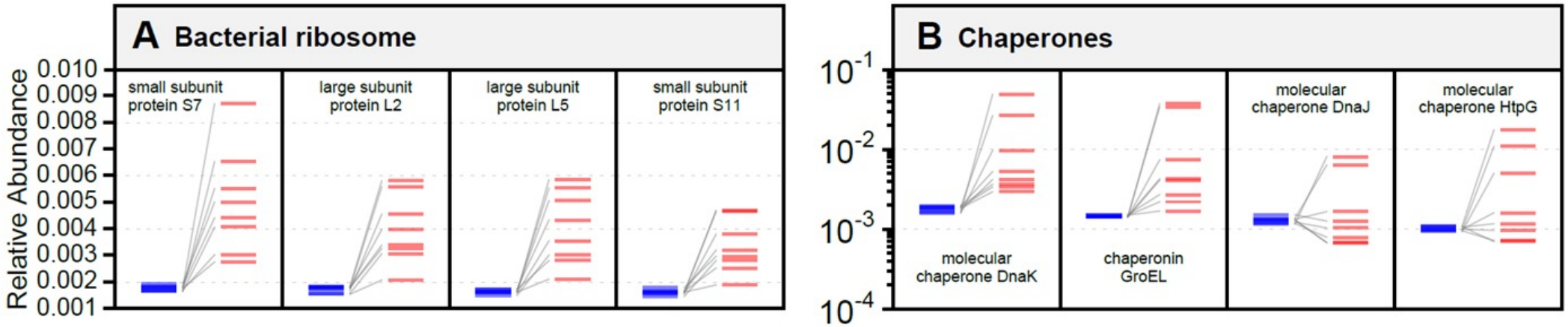
High RNA abundance relative to DNA abundance



Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Activity is more personalized than potential



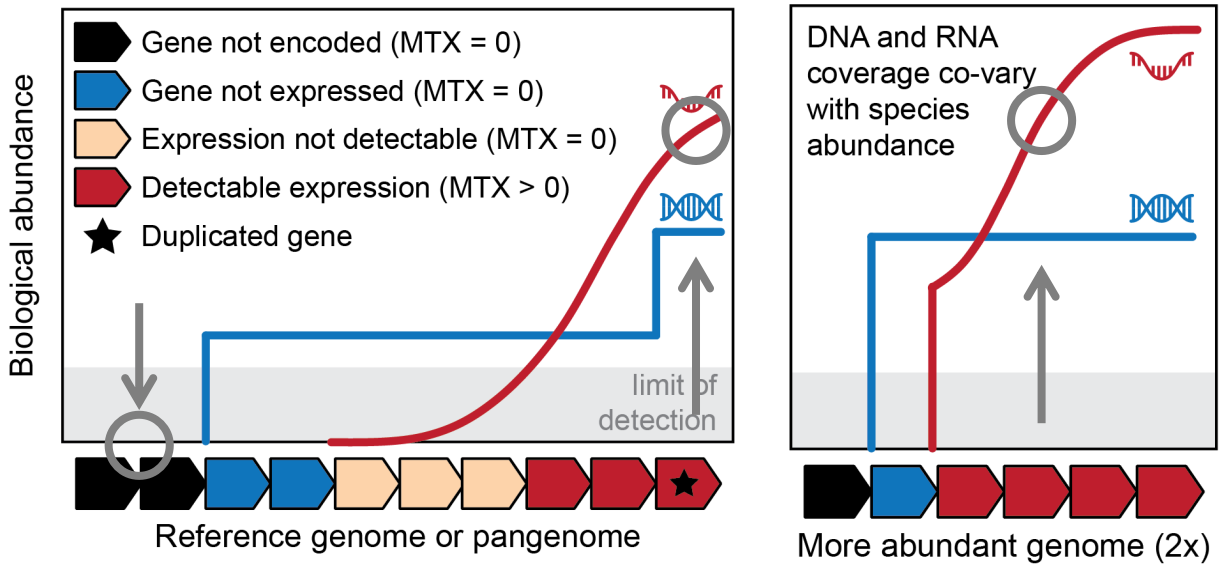


Franzosa et al, *Proc Natl Acad Sci U S A*, 2014

Statistics for metatranscriptomics (MTX)

Yancong
Zhang

Stats challenges in the analysis of MTX data



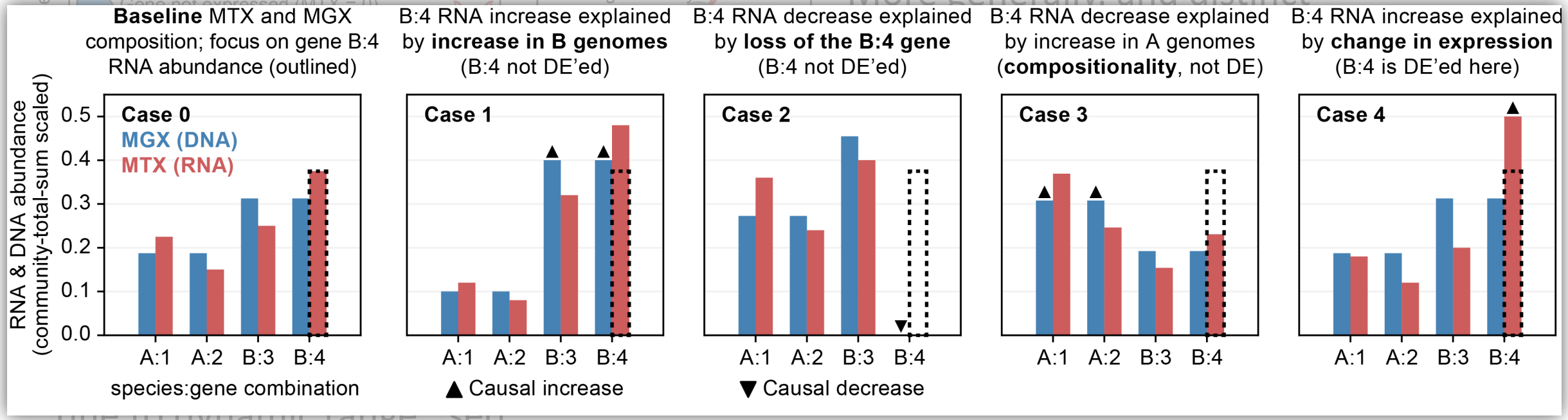
More generally, and distinct from single-species RNA-seq, MTX copy number depends on MGX (gene) copy number

Technical zeros are common due to dynamic range : seq depth interaction

Biological zeros are common from missing species or gene deletion

Zhang et al., *Bioinformatics*, 2021

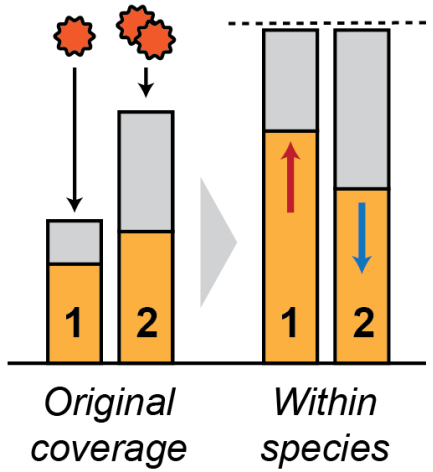
Stats challenges in the analysis of MTX data

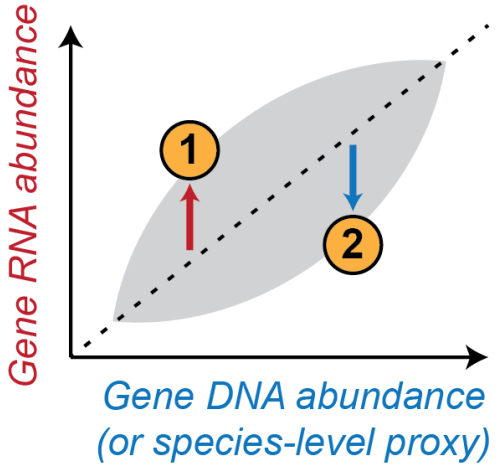


As a consequence, many phenomena that lead to differential abundance in MTX data are not related to differential regulation of microbial genes!

Zhang et al., *Bioinformatics*, 2021

Accounting for “relative expression” (concepts)





Family 1: normalize RNA within each community species
⇒
(*essentially*)
N separate single-species RNA-seq datasets

Family 2: adjust RNA abundance for gene copy number (or a proxy, e.g. species abundance) using a ratio or residual

Zhang et al., *Bioinformatics*, 2021

Six formal models of relative expression…

	Model number: name	Model formulation	Model description
	M1: Naïve RNA	$C(f_{\mathrm{RNA}}) \sim p$	The community total-sum-scaled abundance $C(x)$ of a feature f's RNA count (f_{RNA}) is modeled as a function of a sample phenotype/property (p).
✕	M2: Within-taxon RNA	$T(f_{\mathrm{RNA}}) \sim p$	f_{RNA} is subjected to within-taxon sum-scaling, $T(x)$, i.e. normalizing against the total RNA pool of f's source taxon.
✕	M3: Taxon-RNA covariate	$T(f_{\mathrm{RNA}}) \sim Tax_{\mathrm{RNA}}(f) + p$	f_{RNA} is scaled within-taxon as in M2; an RNA-level estimate of f's source taxon abundance, $Tax_{RNA}(f)$, is included as a covariate.
+	M4: RNA/DNA ratio	$C(f_{\mathrm{RNA}})/C(f_{\mathrm{DNA}}) \sim p$	The ratio of f's total-sum scaled RNA and DNA abundances (the "relative expression ratio") is modeled as a function of p.
✱	M5: Taxon-DNA covariate	$C(f_{\mathrm{RNA}}) \sim Tax_{\mathrm{DNA}}(f) + p$	f_{RNA} is total-sum scaled as in M1; a DNA-level estimate of f's source taxon abundance, $Tax_{DNA}(f)$, is included as a covariate.
+	M6: Feature-DNA covariate	$C(f_{\mathrm{RNA}}) \sim C(f_{\mathrm{DNA}}) + p$	f_{RNA} is total-sum scaled as in M1; a total-sum scaled estimate of f's DNA-level abundance is included as a covariate.

✕ = *The model requires a mapping of features to taxa* + = *the model requires paired MGX (DNA) data* ✱ = *both are required*

Zhang et al., *Bioinformatics*, 2021

…and three modes for pre-filtering technical zeros

Community RNA + DNA for feature f over 6 samples

f_{RNA}	1	0	1	0	2	0
f_{DNA}	0	0	1	0	1	1

Lenient filtering: Treat all zeroes as informative but ignore features that are never seen at the RNA or DNA level.

exclude 2 samples

f_{RNA}	1	0	1	0	2	0
f_{DNA}	0	0	1	0	1	1

Semi-strict filtering: Treat zero values as informative if $\max(f_{\mathrm{RNA}}, f_{\mathrm{DNA}}) > 0$; exclude samples with $\max(f_{\mathrm{RNA}}, f_{\mathrm{DNA}}) = 0$.

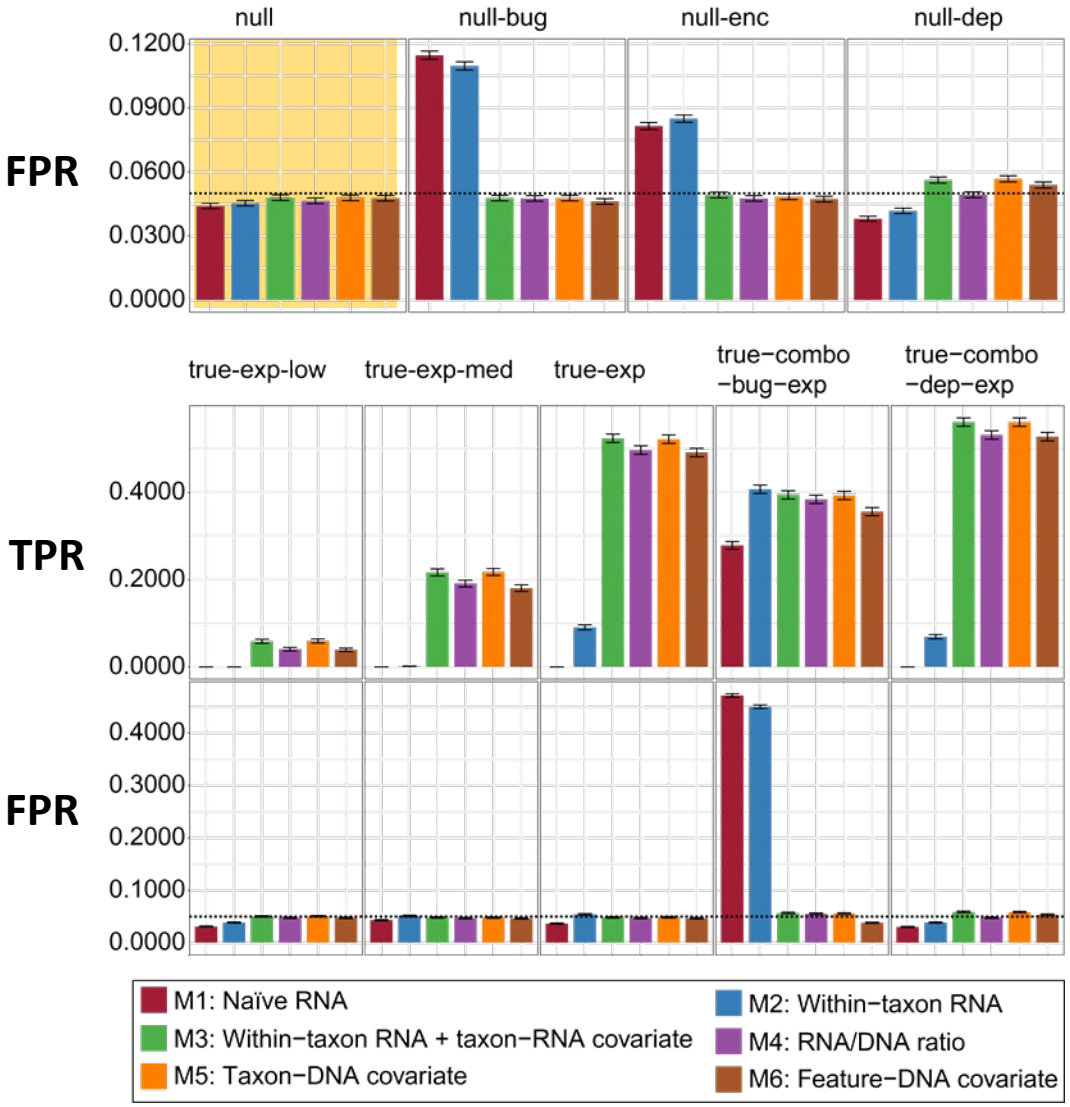
exclude 4 samples

f_{RNA}	1	0	1	0	2	0
f_{DNA}	0	0	1	0	1	1

Strict filtering: Treat all zeroes as likely technical and uninformative; exclude samples with $\min(f_{\mathrm{RNA}}, f_{\mathrm{DNA}}) = 0$.

Zhang et al., *Bioinformatics*, 2021

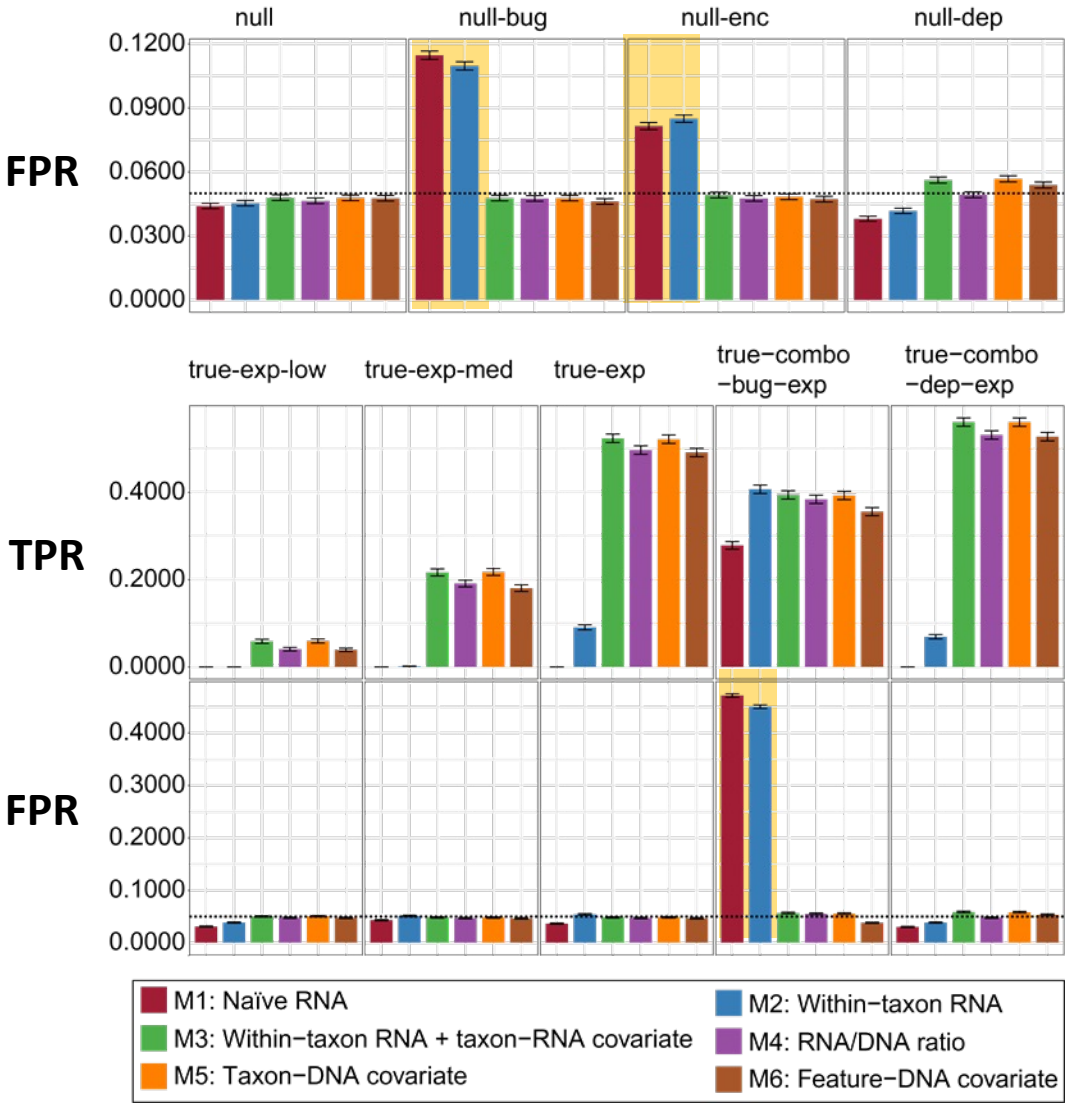
Model performance on synthetic MTX datasets



- (***Based on strict filtering of zeros!***)
- All methods report OK FPR with no signal
- M1 (naïve) and M2 (within-species RNA) confuse bug abundance changes for DE
- M3 (within-species RNA with total RNA covariate) has better FPR and good TPR (no DNA needed)
- DNA-aware methods are roughly equivalent here
- *Note*: Strict filtering of MTX zeros prevents models M1-3 & 5 from confusing gene loss for DE

Zhang et al., *Bioinformatics*, 2021

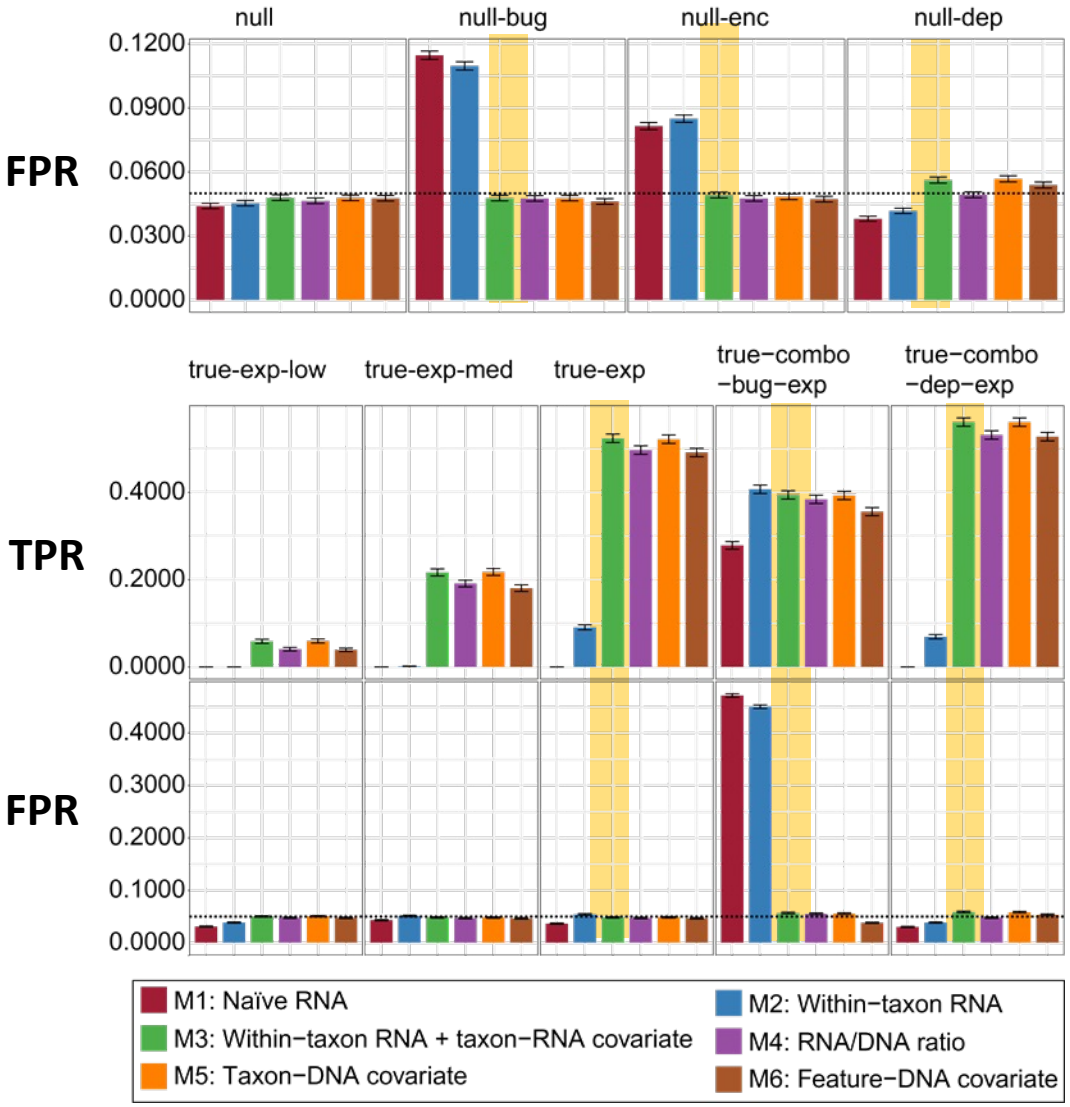
Model performance on synthetic MTX datasets



- (***Based on strict filtering of zeros!***)
- All methods report OK FPR with no signal
- M1 (naïve) and M2 (within-species RNA) confuse bug abundance changes for DE
- M3 (within-species RNA with total RNA covariate) has better FPR and good TPR (no DNA needed)
- DNA-aware methods are roughly equivalent here
- *Note*: Strict filtering of MTX zeros prevents models M1-3 & 5 from confusing gene loss for DE

Zhang et al., *Bioinformatics*, 2021

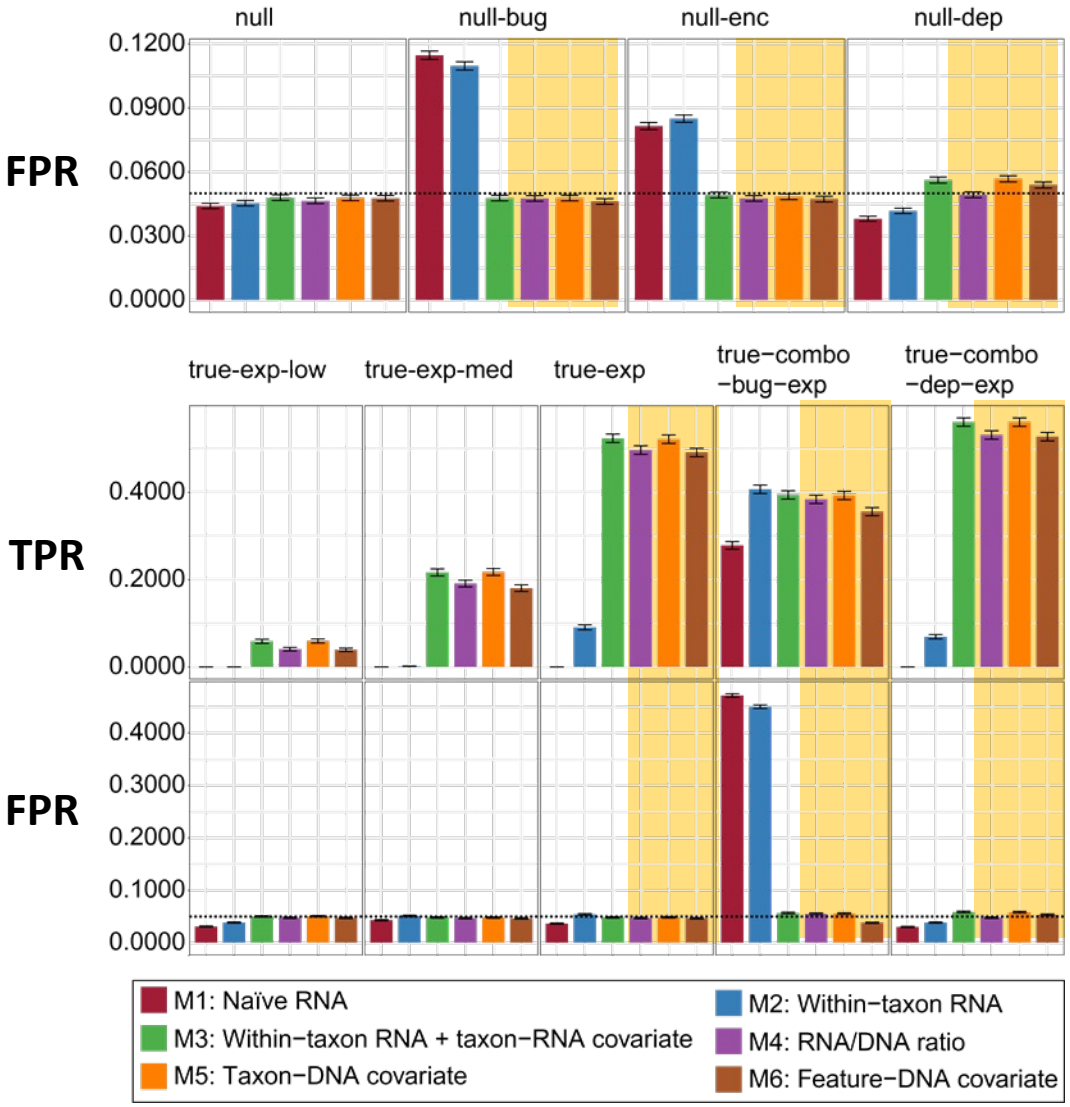
Model performance on synthetic MTX datasets



- (***Based on strict filtering of zeros!***)
- All methods report OK FPR with no signal
- M1 (naïve) and M2 (within-species RNA) confuse bug abundance changes for DE
- M3 (within-species RNA with total RNA covariate) has better FPR and good TPR (no DNA needed)
- DNA-aware methods are roughly equivalent here
- *Note*: Strict filtering of MTX zeros prevents models M1-3 & 5 from confusing gene loss for DE

Zhang et al., *Bioinformatics*, 2021

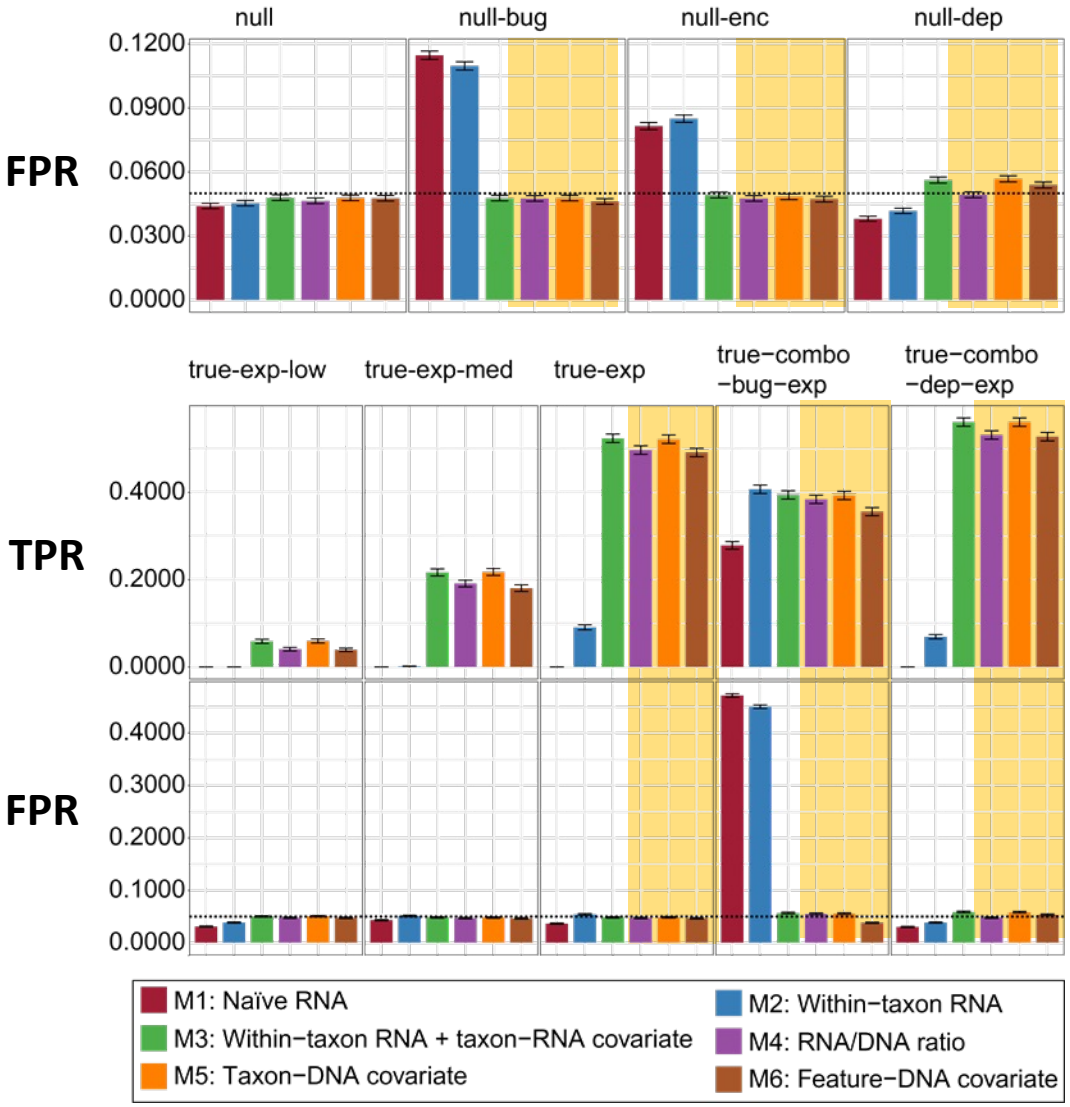
Model performance on synthetic MTX datasets



- (***Based on strict filtering of zeros!***)
- All methods report OK FPR with no signal
- M1 (naïve) and M2 (within-species RNA) confuse bug abundance changes for DE
- M3 (within-species RNA with total RNA covariate) has better FPR and good TPR (no DNA needed)
- DNA-aware methods are roughly equivalent here
- *Note*: Strict filtering of MTX zeros prevents models M1-3 & 5 from confusing gene loss for DE

Zhang et al., *Bioinformatics*, 2021

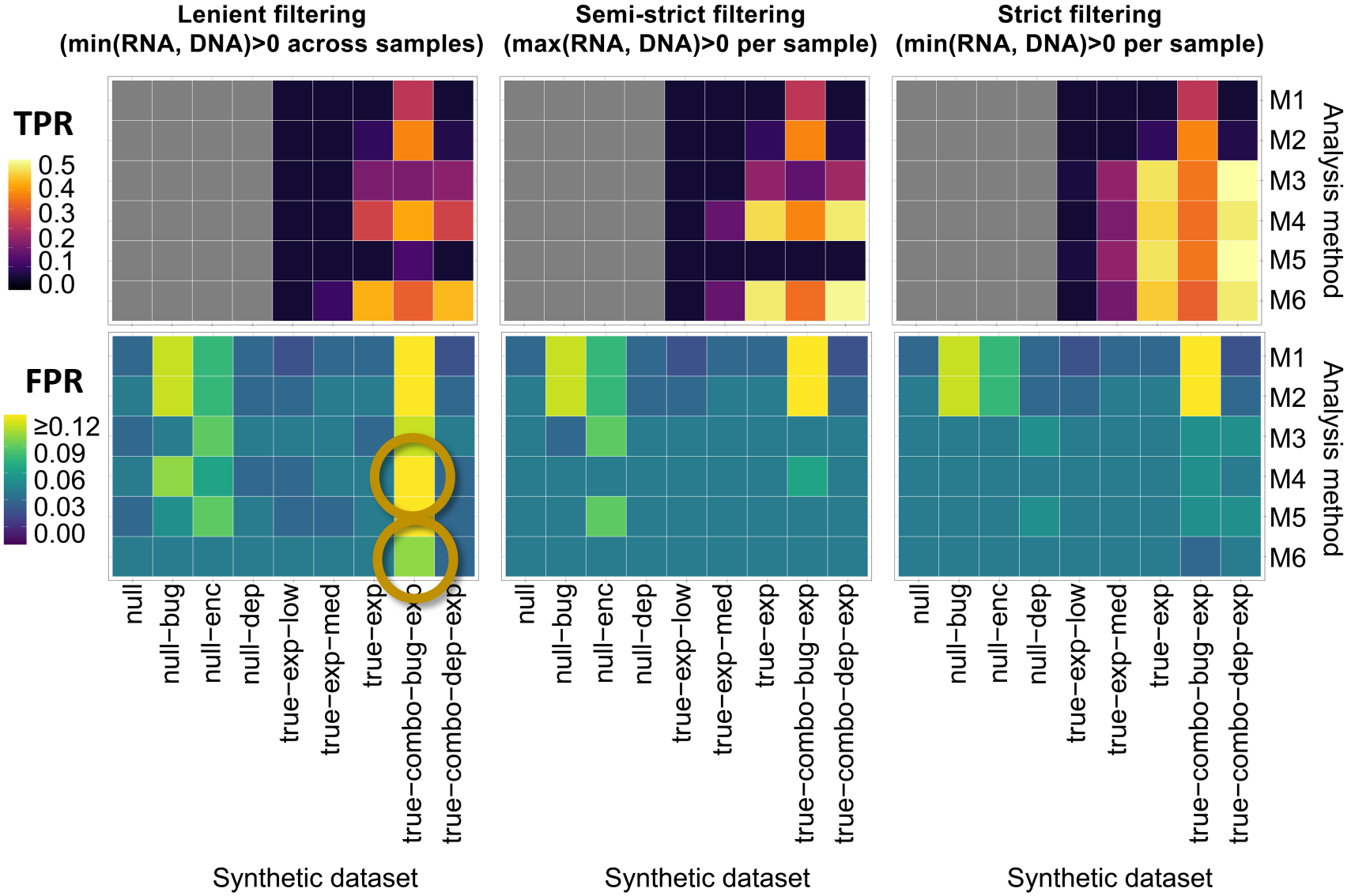
Model performance on synthetic MTX datasets



- (***Based on strict filtering of zeros!***)
- All methods report OK FPR with no signal
- M1 (naïve) and M2 (within-species RNA) confuse bug abundance changes for DE
- M3 (within-species RNA with total RNA covariate) has better FPR and good TPR (no DNA needed)
- DNA-aware methods are roughly equivalent here
- *Note*: Strict filtering of MTX zeros prevents models M1-3 & 5 from confusing gene loss for DE

Zhang et al., *Bioinformatics*, 2021

Model performance vs. zero-filtering method



- These are the results from the previous slide.
- M3 (best no-DNA model) struggles in the presence of technical zeros AND calls gene loss as DE.
- M5 (normalizing by species DNA abundance) behaves similarly.
- M6 (gene DNA covariate) is more robust to zeros than M4 (RNA/DNA ratio) given strong confounding.

Zhang et al., *Bioinformatics*, 2021

Current recommendations for MTX stats

	Model number: name	Model formulation
	M1: Naïve RNA	$C(f_{\mathrm{RNA}}) \sim p$
✕	M2: Within-taxon RNA	$T(f_{\mathrm{RNA}}) \sim p$
✕	M3: Taxon-RNA covariate	$T(f_{\mathrm{RNA}}) \sim Tax_{\mathrm{RNA}}(f) + p$
+	M4: RNA/DNA ratio	$C(f_{\mathrm{RNA}})/C(f_{\mathrm{DNA}}) \sim p$
✱	M5: Taxon-DNA covariate	$C(f_{\mathrm{RNA}}) \sim Tax_{\mathrm{DNA}}(f) + p$
+	M6: Feature-DNA covariate	$C(f_{\mathrm{RNA}}) \sim C(f_{\mathrm{DNA}}) + p$

If you **don't** have paired MGX data, M3 is a reasonable approach if…

- You can assign your transcripts to species
- You ignore all zero values

If you **do** have paired MGX data, M4 and both M6 behave well, with M6 (the gene-DNA covariate) proving more robust to zero values than the RNA/DNA ratio.

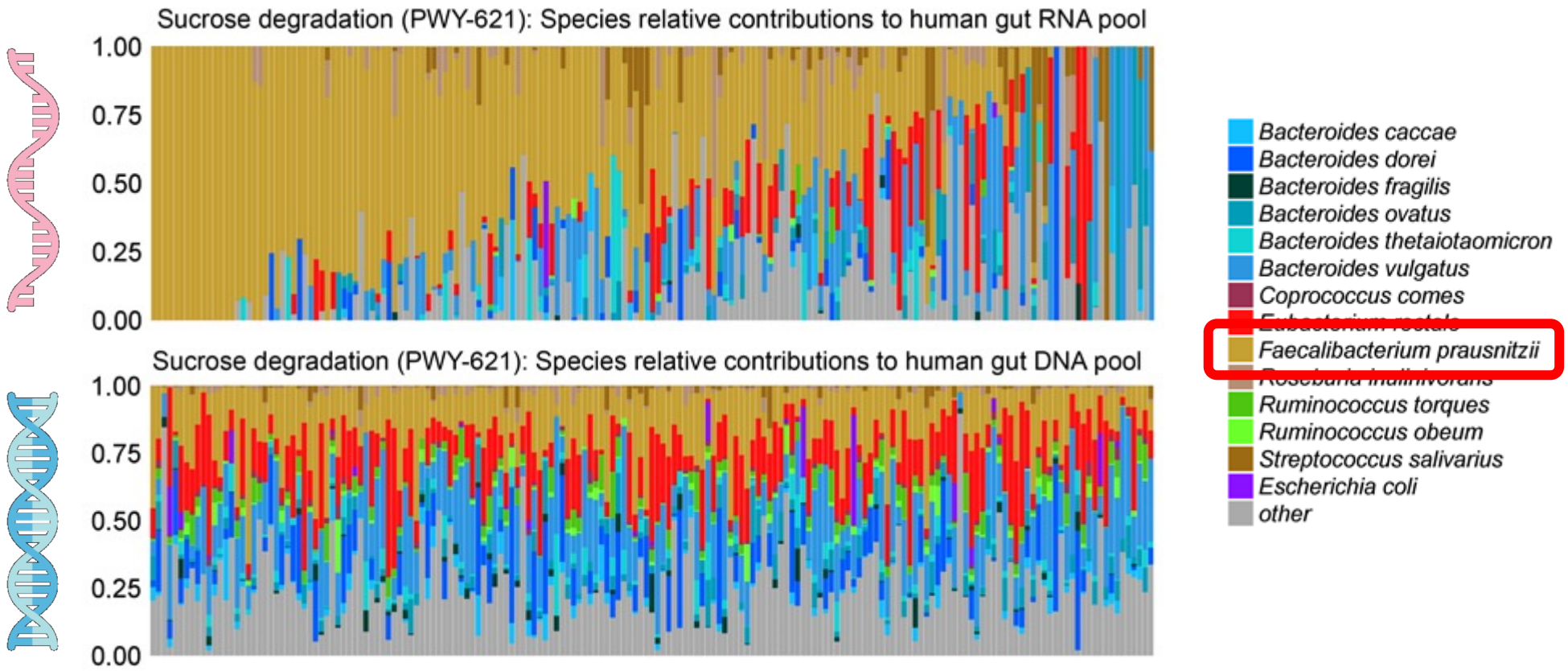
M6 is our current default.

✕ = *The model requires a mapping of features to taxa* + = *the model requires paired MGX (DNA) data* ✱ = *both are required*

Zhang et al., *Bioinformatics*, 2021

Aside: Contributional changes in activity

Abu-Ali et al., *Nat Microbiol*, 2017



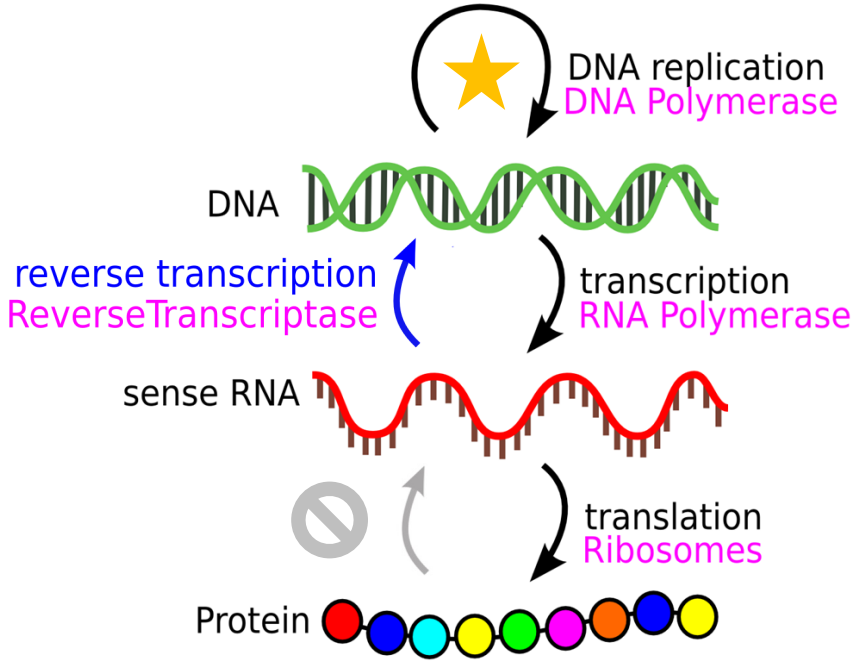
In collaboration with the STARR Consortium & HPFS cohort

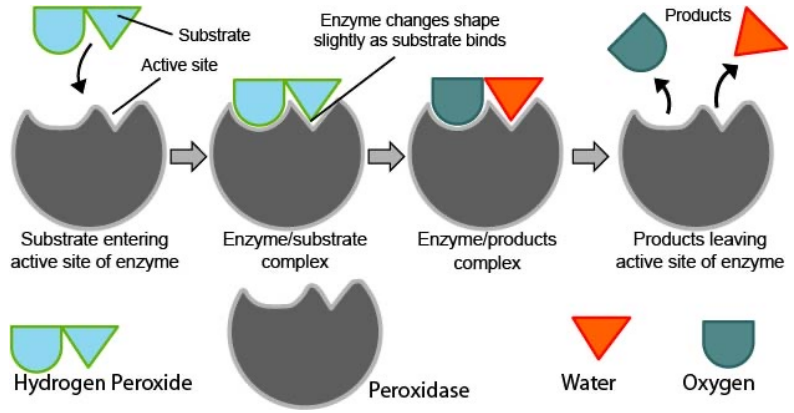
Sucrose degradation follows a complex attribution pattern across ~200 human gut metagenomes…

…but its expression can be dominated by a single species in paired gut metatranscriptomes!

Metabolomics

Why metabolomics and proteomics are different

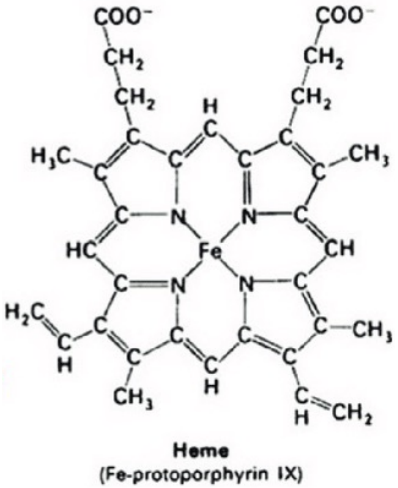


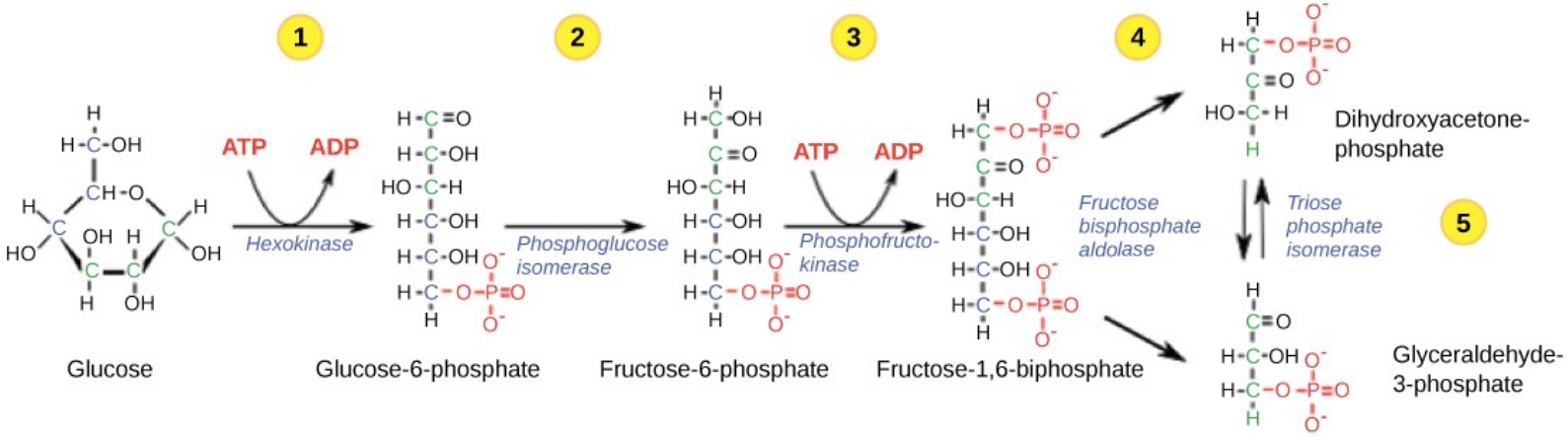


- DNA "wants" to replicate
  - Most sequencing is sequencing by synthesis
- Many problems solved directly by DNA sequencing
  - Genomics, metagenomic profiling
- Other 'omics problems solved by reducing to DNA seq.
  - Transcriptomics (reverse-transcribe RNA to DNA, then seq.)
  - Chromatin conformation (split and rejoin local DNA, then seq.)
  - Protein interaction (seq. interacting genes in rescued cells)
- No biological "reverse translation" (protein → DNA)
  - Quantifying proteins (proteomics) needs another approach
  - (Aside: we can and will map protein to DNA with e.g. BLAST)
- Metabolites aren't a part of the central dogma
  - (Though they *are* acted on by proteins)
  - Metabolomics will require different approaches
  - *We'll start here…*

What is metabolomics?

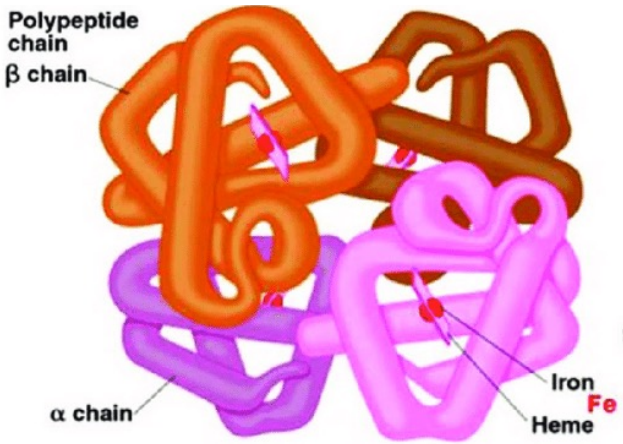
- The study of small molecules (metabolites) produced by a biological system
  - Small usually defined as <1 kda
  - Mass of glucose ~ 0.18 kda
  - Mass of heme ~ 0.62 kda
  - Mass of hemoglobin alpha ~ 16 kda
- Sometimes distinguished from *metabonomics*
  - Includes other molecules in the system, e.g. food
  - Term more common in NMR literature



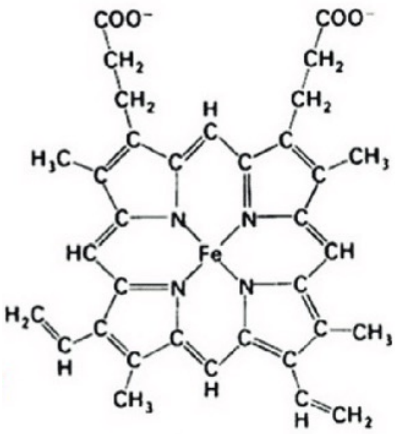


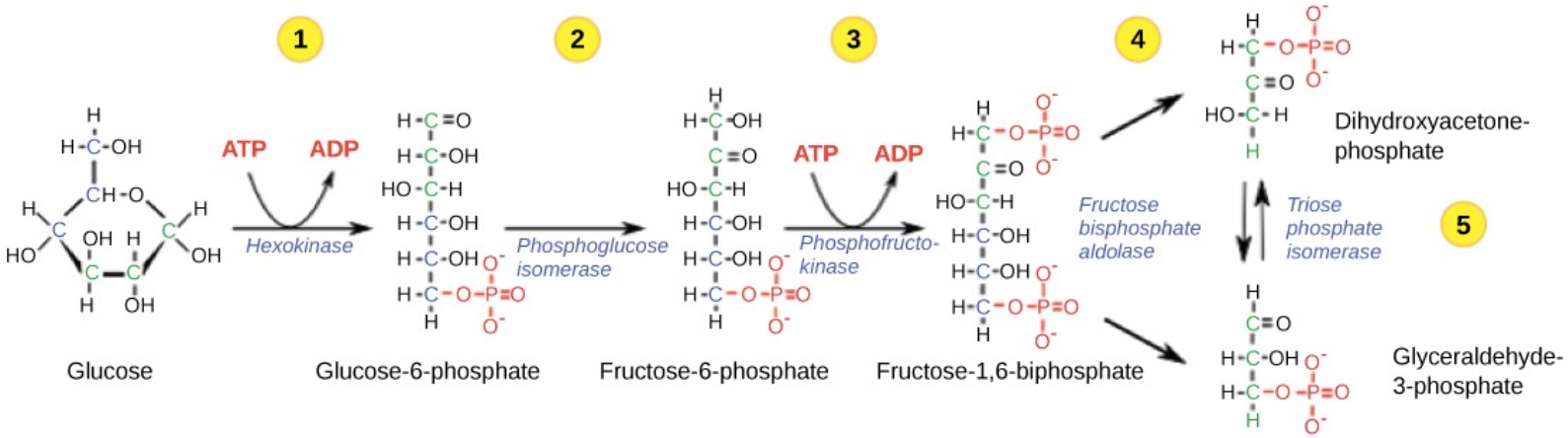
What is metabolomics?

- The study of small molecules (metabolites) produced by a biological system
  - Small usually defined as <1 kda
  - Mass of glucose ~ 0.18 kda
  - Mass of heme ~ 0.62 kda
  - Mass of hemoglobin alpha ~ 16 kda
- Sometimes distinguished from *metabonomics*
  - Includes other molecules in the system, e.g. food
  - Term more common in NMR literature



Hemoglobin


Heme
(Fe-protoporphyrin IX)



Metabolomics philosophies

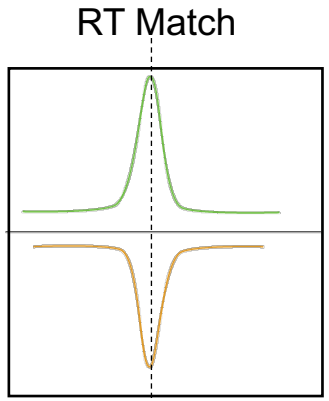
- **<u>Targeted</u> metabolomics**
  - Quantifies 10s of well characterized metabolites from a sample
  - Provides absolute quantification through comparisons with chemical standards
  - “Usual suspects”: central carbon metabolism, amino acids, nucleotides
- **<u>Untargeted</u> metabolomics**
  - Attempts to enumerate 1,000s of unique metabolites from a sample
    - Only characterizing what you can, and with mixed accuracy
  - Usually requires multiple screens targeting different chemistries
  - Provides (mostly) relative abundances

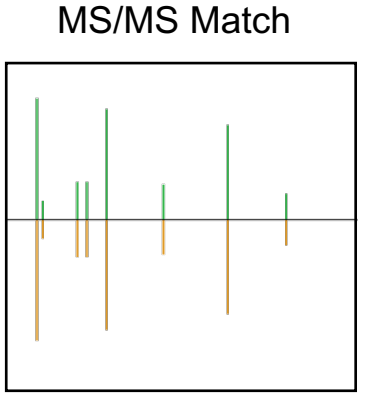
Metabolomics technologies

- **Nuclear Magnetic Resonance (NMR) spectroscopy**
  - Uses environment-specific atomic resonance (e.g. ^{1}H) patterns to identify compounds
  - Can also be used to work out molecular structures (from coupling)
  - *Pros*: Absolute, reproducible quantification; minimal sample prep; sample unperturbed
  - *Cons*: Low sensitivity (# of molecules, min. abundance); machines very expensive
- **Mass spectrometry (MS)**
  - Directly measures mass/charge ratio of a metabolite
  - Can also be used to work out molecular structure (from fragmentation pattern)
  - *Pros*: Very sensitive and flexible; can differentiate 1,000s of metabolites
  - *Cons*: Relative quantification only; sample destroyed; complex preps

Metabolite identification

- Work out mass (or more properly mass-to-charge ratio, m/z) and thus potential formulas for parent compound
- Match against metabolite databases (e.g. HMDB, Metlin)
  - Likely to be lots of hits
- Match against compound libraries
  - Internal RT values, known/theoretical MS^2 fragmentation

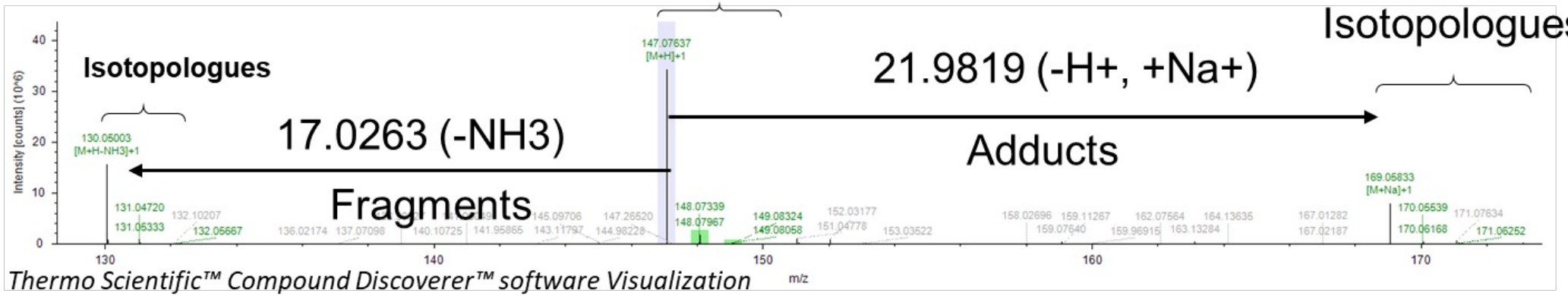




At a given retention time:

$[M+H]^+$ $C_5H_{10}N_2O_3$

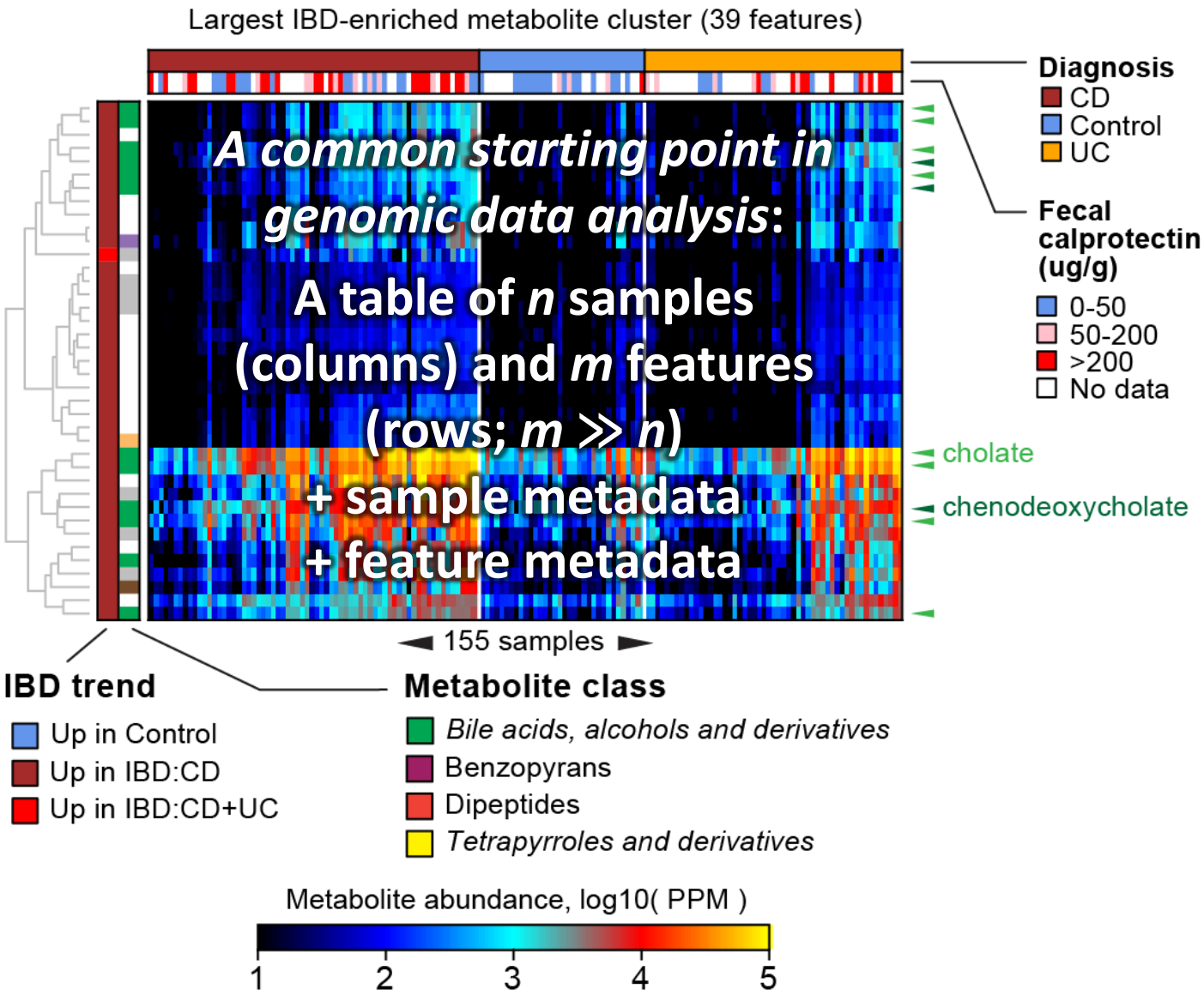
Isotopologues



Thermo Scientific™ Compound Discoverer™ software Visualization

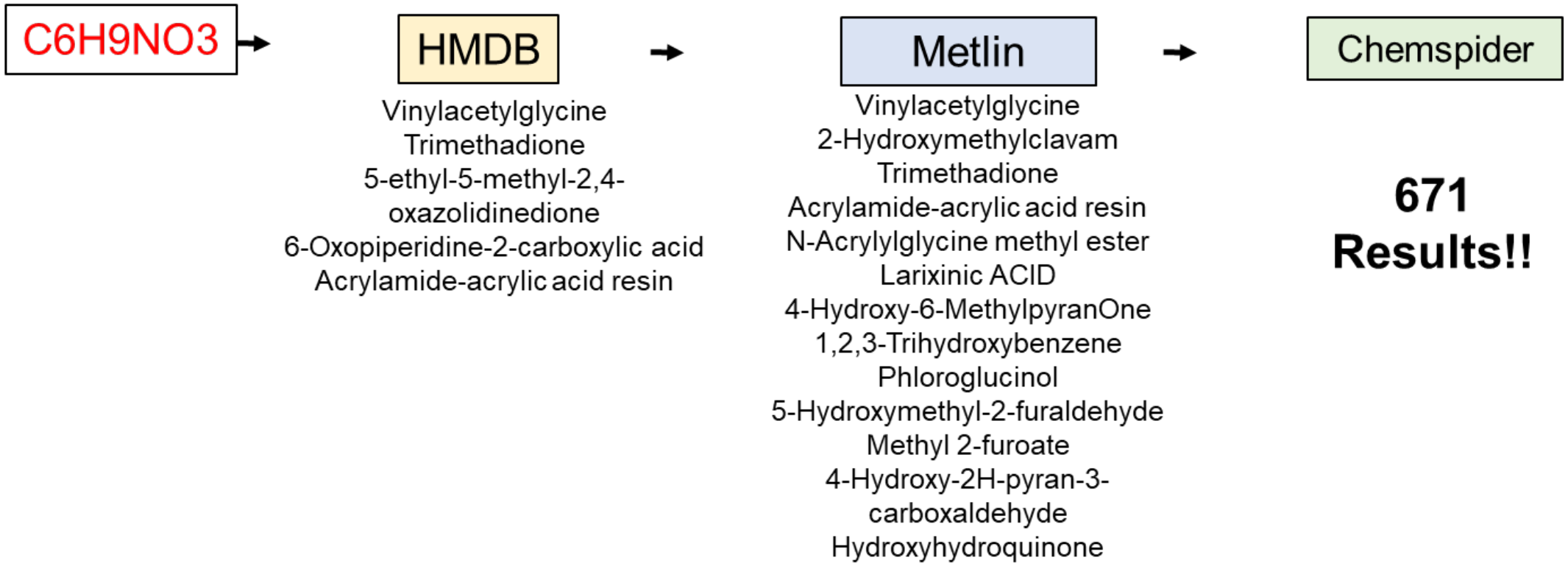
What to do with your sample-by-metabolite table?

- Quality control
  - Normalization, outlier removal
- Supervised analyses
  - Phenotype association/prediction
  - Enrichment analyses
- Unsupervised analyses
  - Clustering samples
  - Clustering metabolites
- Data integration
  - Associating metabolites with other 'omics measurements
  - Do characterized metabolites associate with uncharacterized genes?
  - Do uncharacterized metabolites associate with characterized genes?



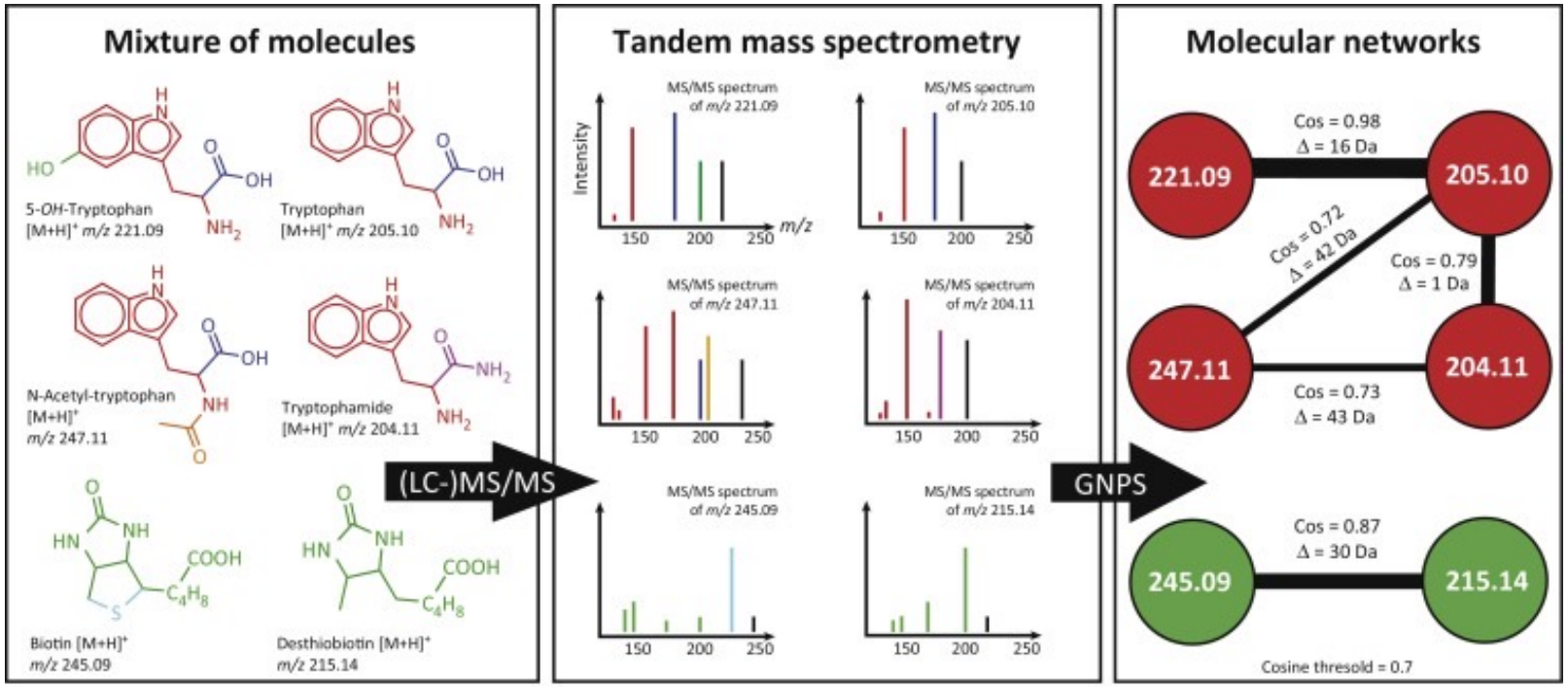
Metabolite identification

- A formula alone can be very ambiguous (even worse if allowing ±1 da)



Metabolite identification

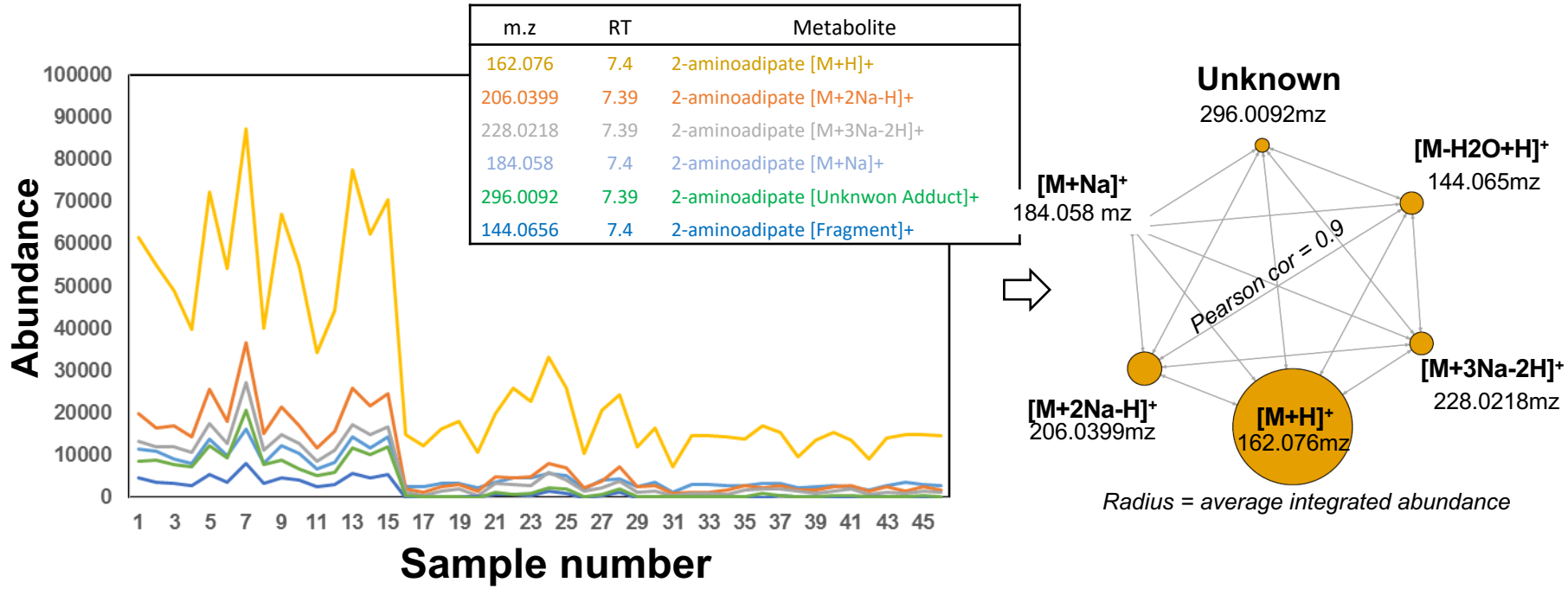
- Analyze networks of similar fragmentation spectra (MS/MS only)
  - “Molecular Networking” by Pieter Dorrestein / GNPS
- Transfer knowledge from knowns to unknowns (guilt-by-association)



Trends in Pharmacological Sciences

Metabolite identification

- Analyze cross-sample covariation
  - Identify groups of related metabolites (similar to co-expression)
  - Transfer annotations from known to unknown metabolites (also guilt-by-association)



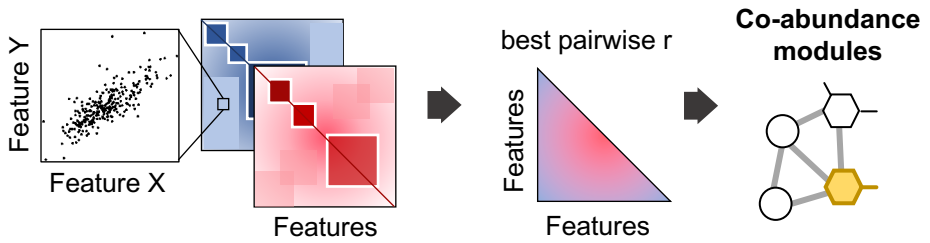
Brown, et al (2011) Bioinformatics 27: 1108−1112

Parent ion
+ isotopes
+ adducts
+ common fragments
(all in same RT group)

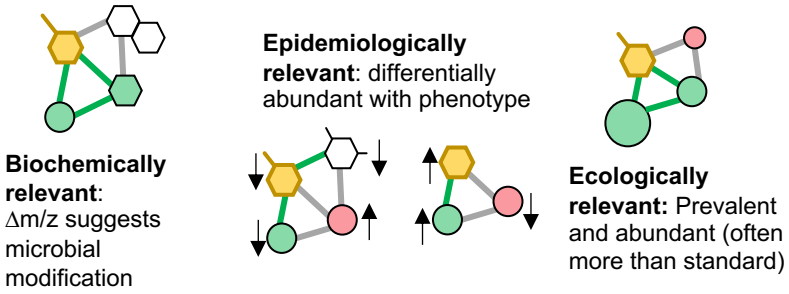
MACARRoN: Metabolome Analysis and Combined Annotation Ranks for pRediction of Novel bioactives

Amrisha Bhosle

Guilt-by-association to transfer potential annotations



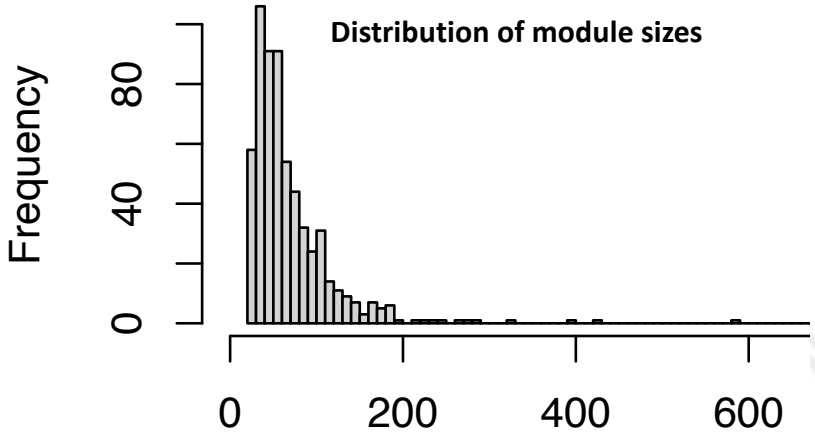
Signs of potential bioactivity



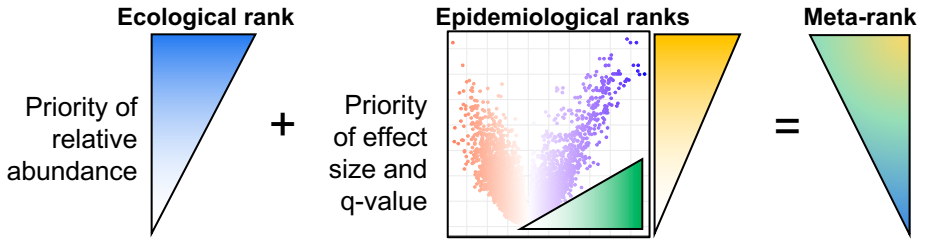
546 HMP2 fecal metabolomes
(non-IBD + CD + UC)

- 44,757 features
- **466 annotated**

43,498 features in 606 modules



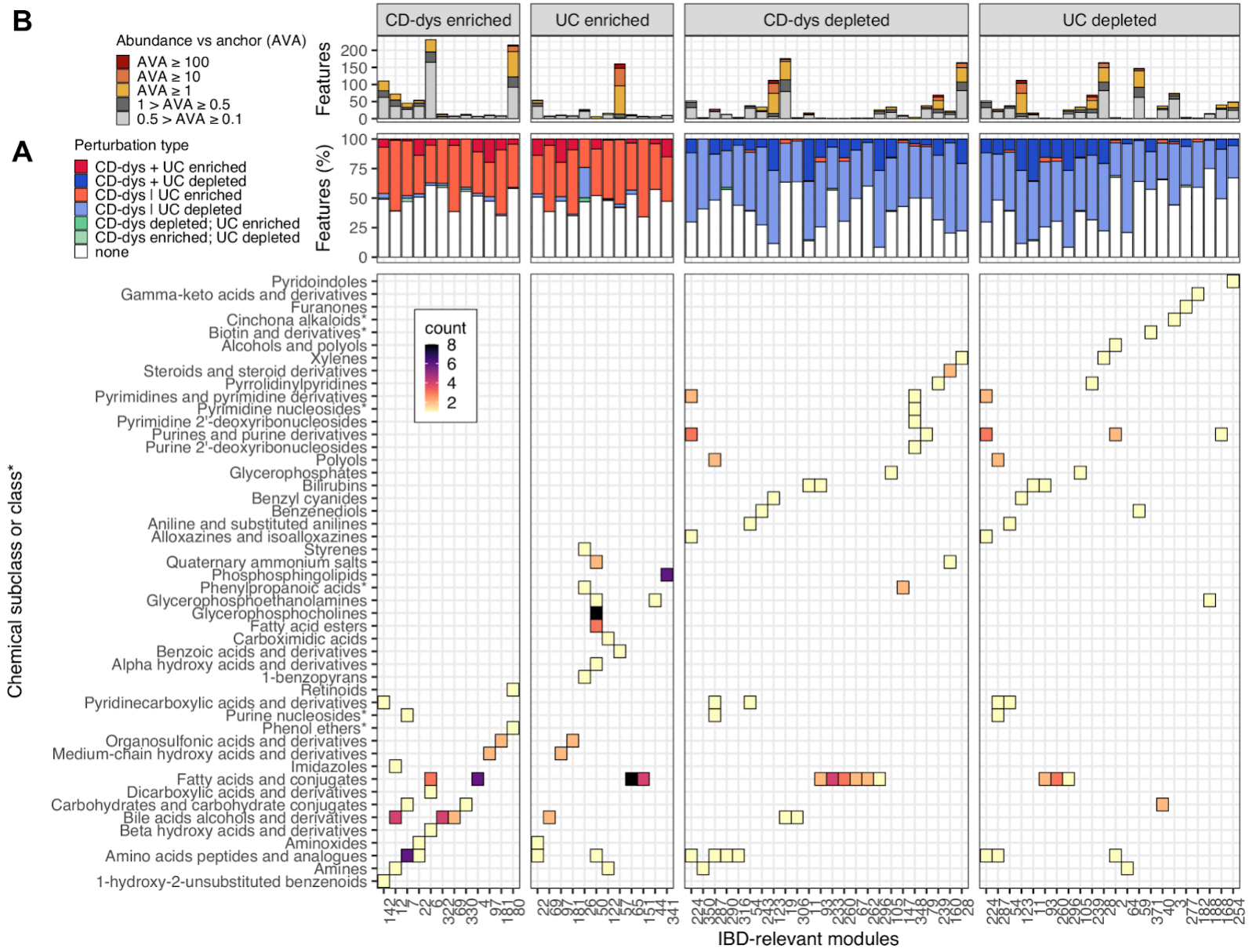
Integrated prioritization of potential bioactives



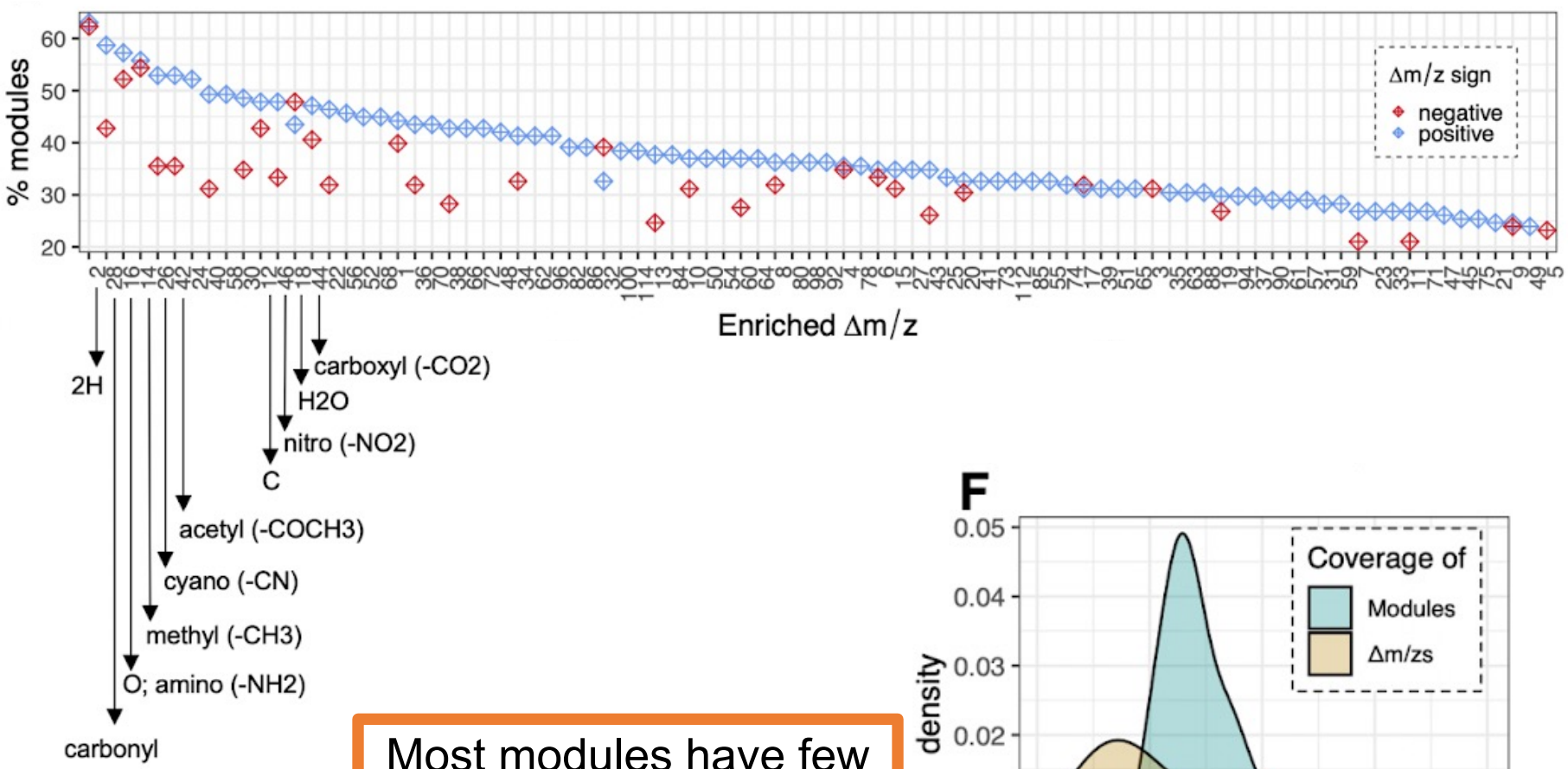
152 modules with at least one annotated metabolite→
13,633 features associated with at-least one annotated metabolite

http://huttenhower.sph.harvard.edu/macarron

A large subset of modules are IBD-relevant and characterizable

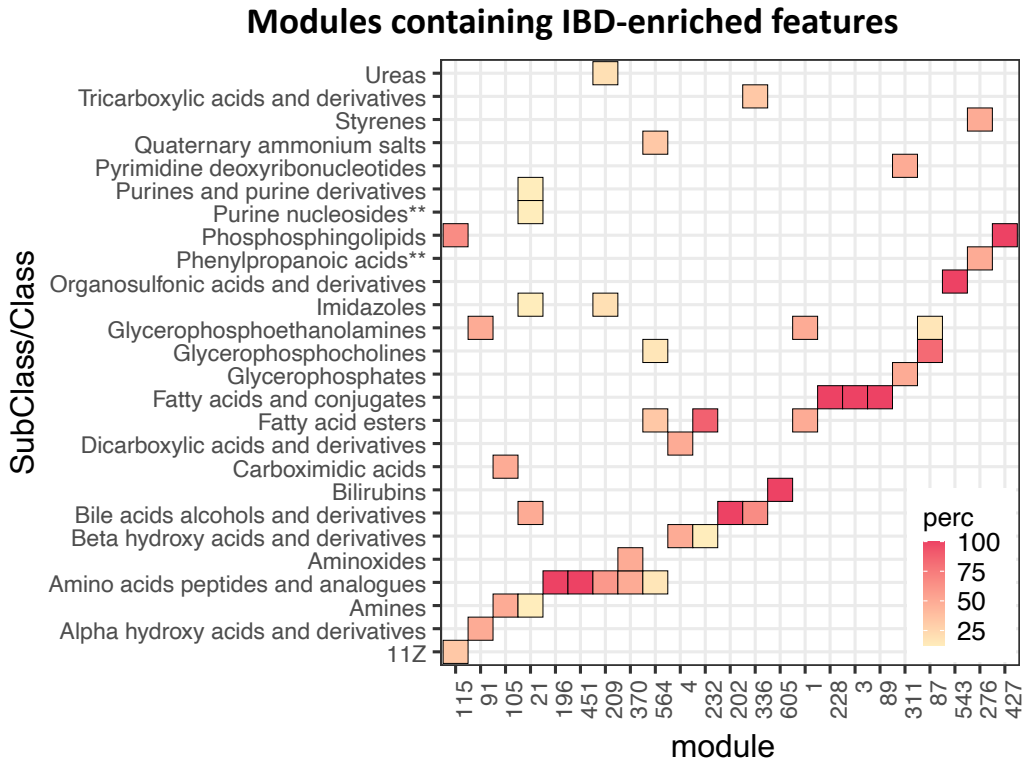


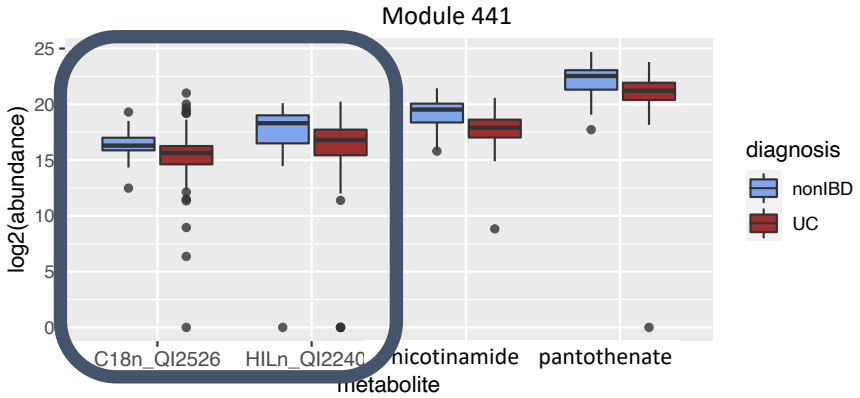
Putative modifications are characteristic of specific parent compounds



Most modules have few Δm/zs, but most Δm/zs appear somewhere.

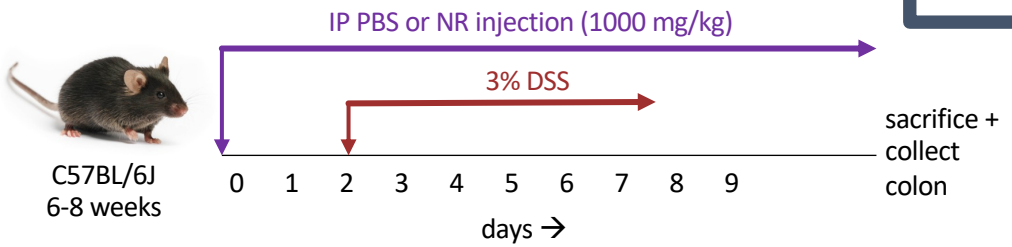
Bioactivity predictions are consistent and (at least one is) validated *in vivo*



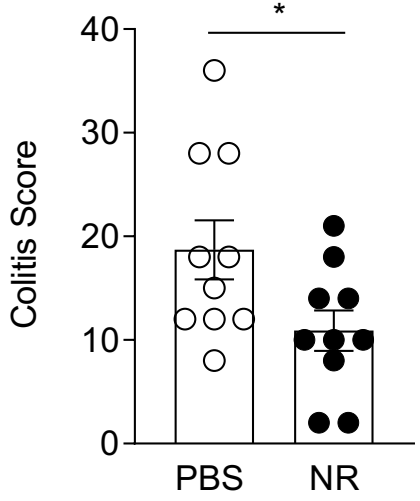


- Two prioritized metabolites depleted in UC.
- Unidentified in HMP2 dataset.
- Inferred as nicotinamide riboside (B3).

Dextran sulfate sodium (DSS) model for injury / colitis



Nicotinamide riboside ameliorates gut injury and inflammation.




Wendy Garrett


Sena Bae

Summary

- Metatranscriptomics (MTX)
  - Functional profiling of MTX data
  - Relationships between functional potential and activity
  - Detecting under- and over-expression
  - Statistical models for MTX analysis
- Metabolomics (MBX)
  - Sequencing is easy… everything else is hard
  - Metabolomics approaches
  - Identifying compounds
  - Bioactivity prediction with MACARRoN

(One way of thinking about)

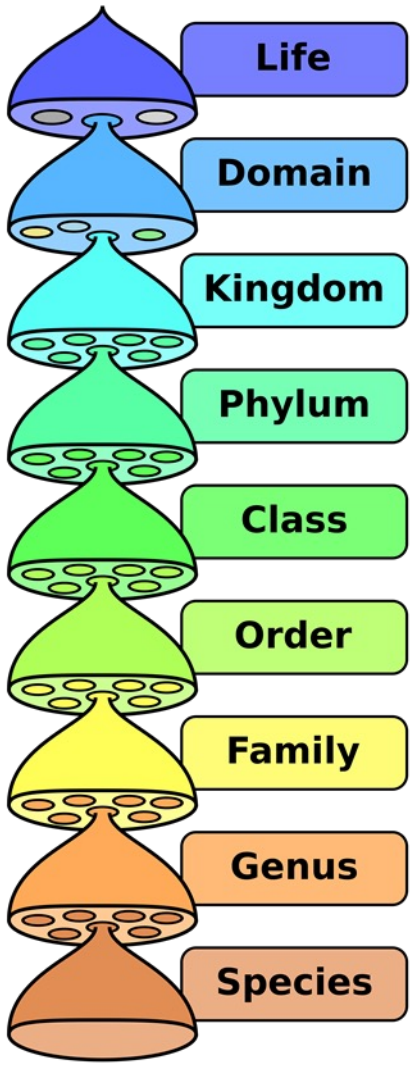
Strains in microbial communities

Summary

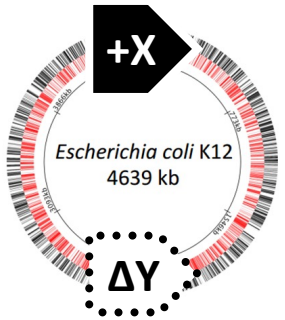
- What is a microbial strain?
- Why is considering strains important?
- Gene-level strain profiling with PanPhlAn
- SNV-level strain profiling with StrainPhlAn
- Community strain statistics with ANPAN

Two big questions of shotgun meta'omics

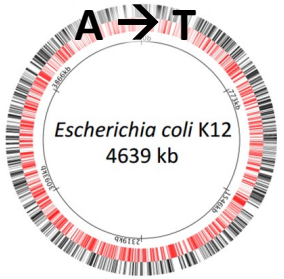
What is a microbial strain?



- A microbial strain is a taxonomic group below the species level
  - Defining species is not trivial, though systematic definitions exist
  - 95% average nucleotide identity; 97-99% 16S rRNA identity
- Strain = population of clonal cells
  - OK for culture applications, but less applicable to sequencing
- Strain = specific microbial isolate genome sequence
  - Very precise, but often too specific to be useful
- Strain = collection of “very similar” cells/genomes
  - With “very similar” defined using phylogenetic criteria



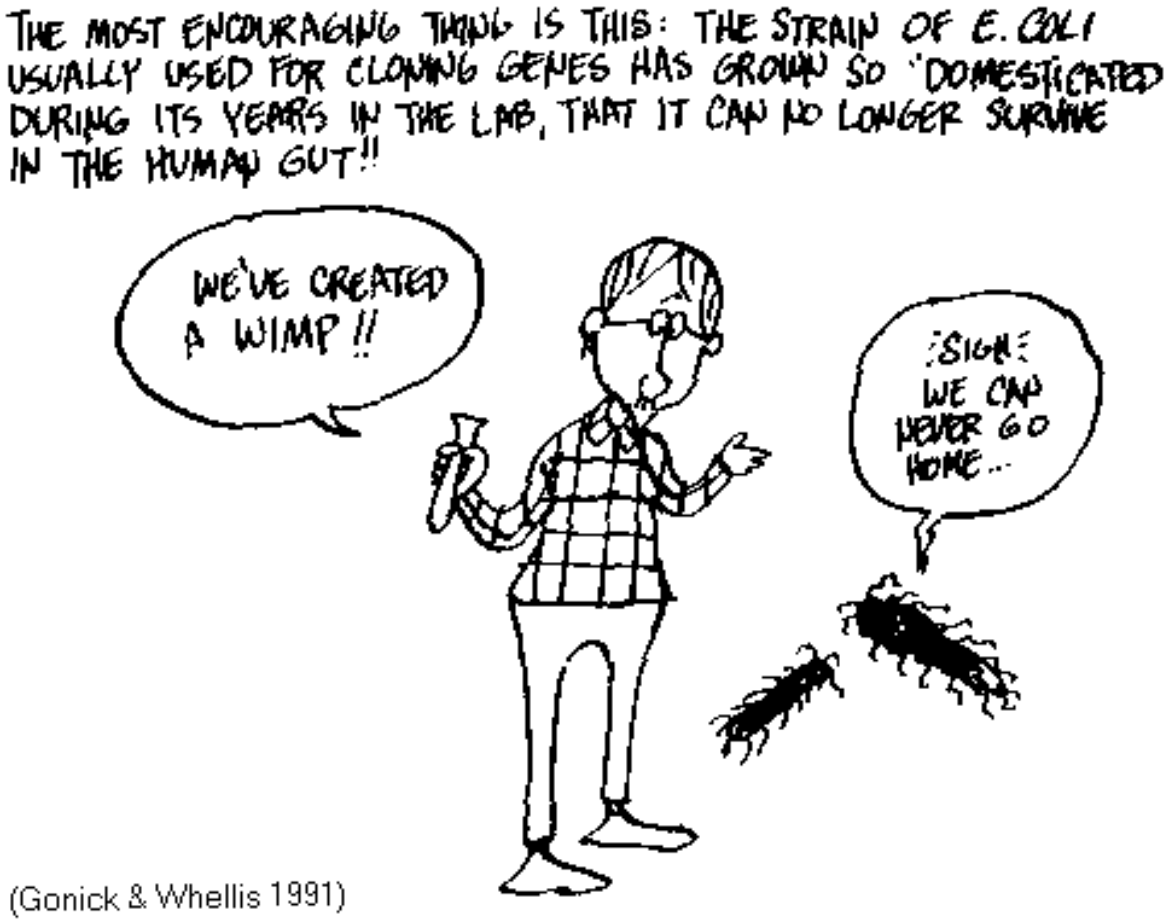
Does adding a gene to a species’ genome make a new strain?
How about deleting one?



What about changing a single nucleotide?

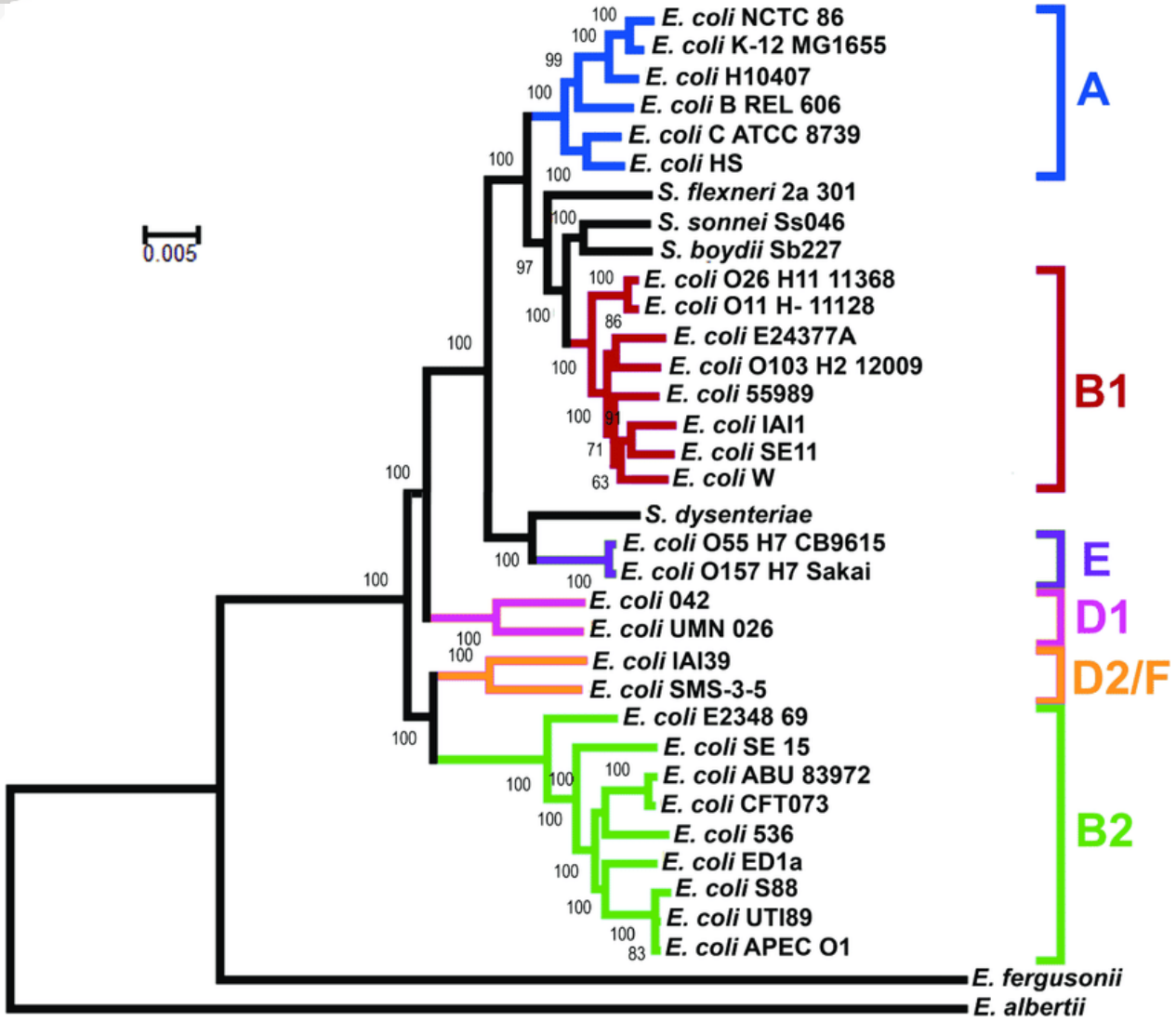
What about 10 or 100?

Why care about strains? Consider *E. coli*



(Gonick & Whellis 1991)

Why care about strains? Consider *E. coli*



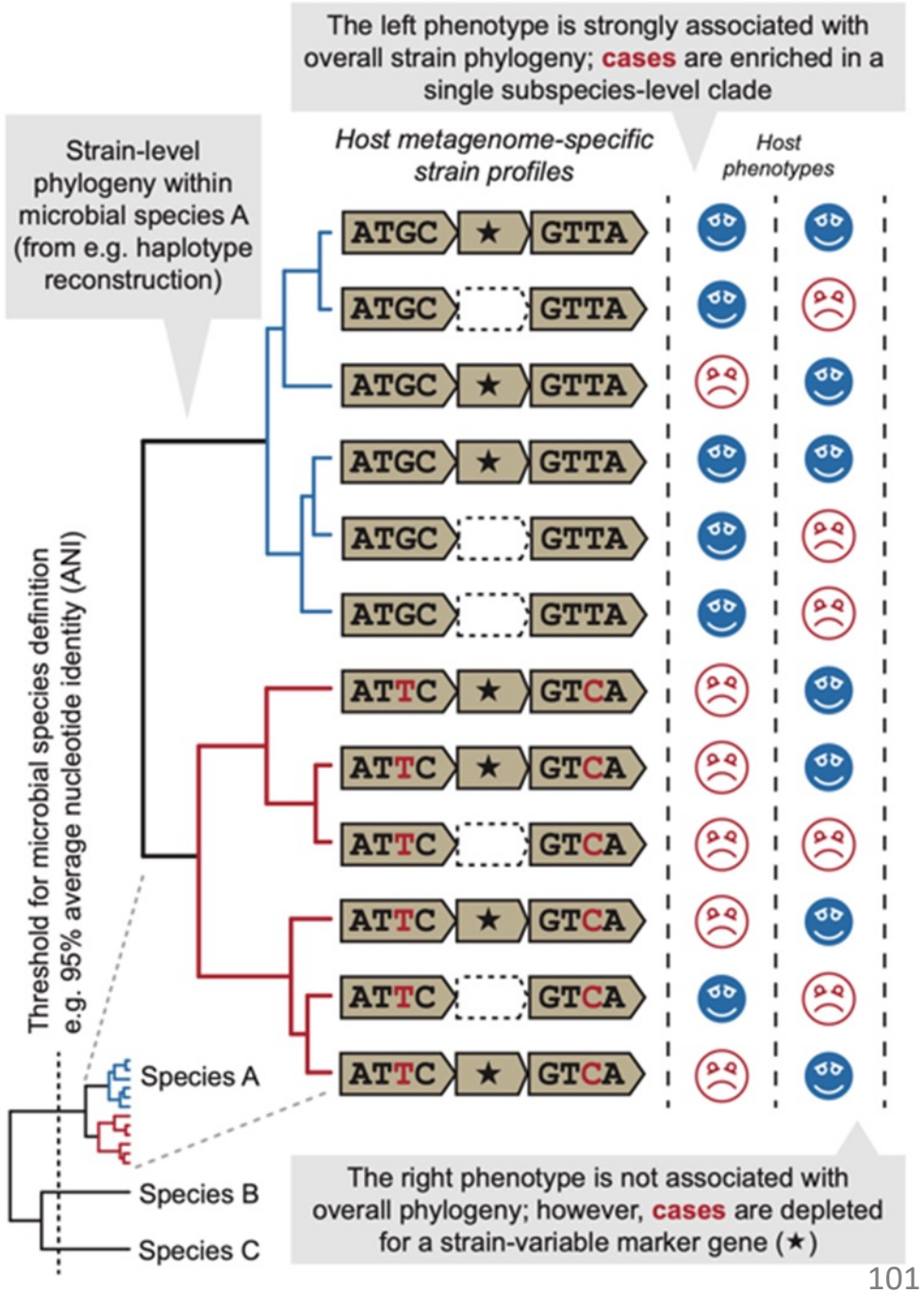
Dunne, Microbial Genomics, 2017

- 1,000s of genomes sequenced
  - Clear sub-species structure
- Some strains are benign
  - Common in the human gut
- Some are acute pathogens
  - e.g. *E. coli* O157:H7
- Some are risk-associated
  - e.g. pks+ *E. coli* in colon cancer
- Some are probiotic
  - *e.g. E. coli* Nissle 1917

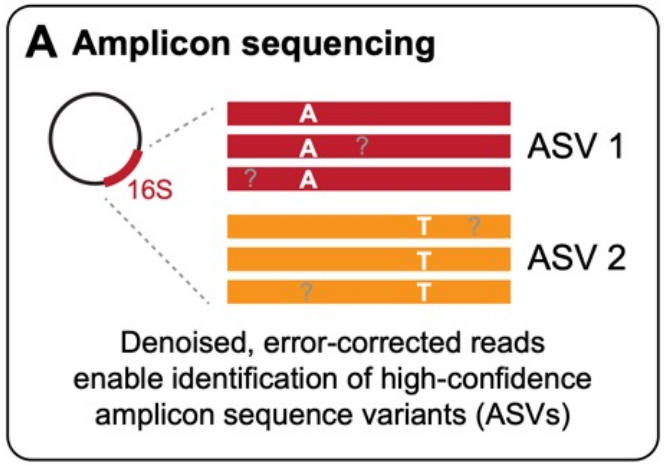
What if we looked at other microbial species in the same detail as E. coli?

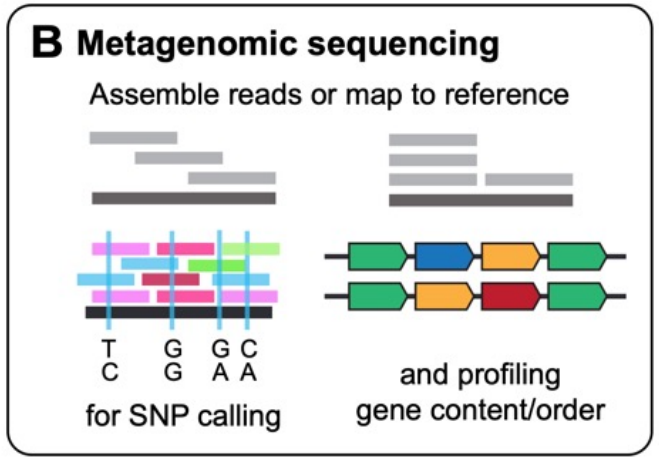
Why care about strains?

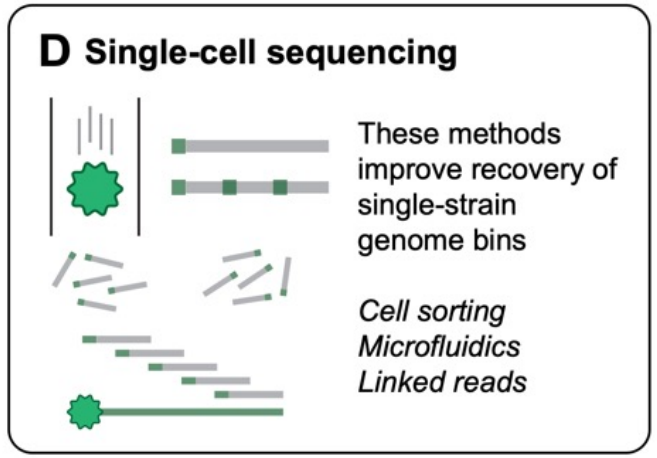
- Strain-level variation can associate with a phenotype or environmental property even when species do not.
- A phenotype ☹ can associate with species phylogeny below the species level, or…
- …a phenotype ☹ can associate with carriage of a specific gene.
- (*Note that neither of these associations requires defining clear sub-species*.)

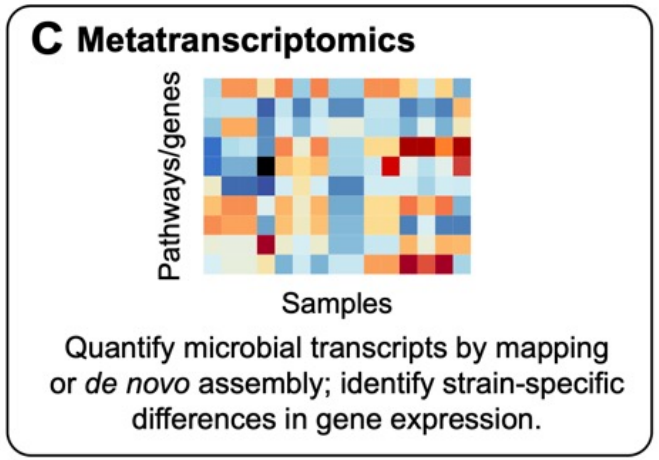


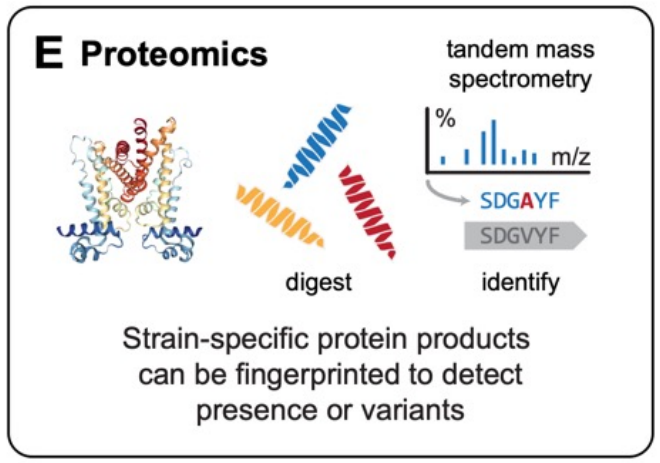
How to study strains in a culture-free context?

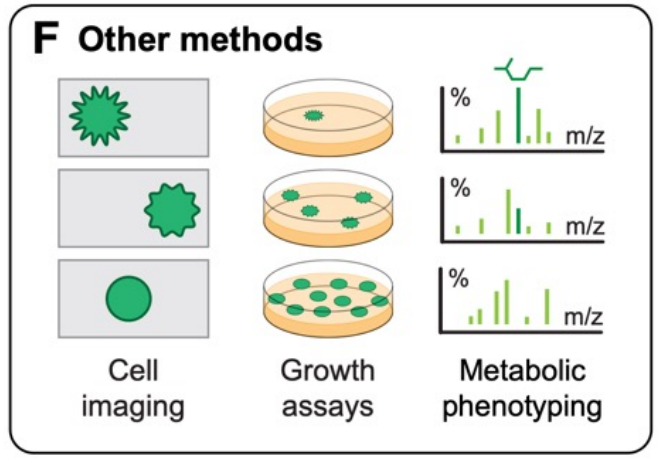




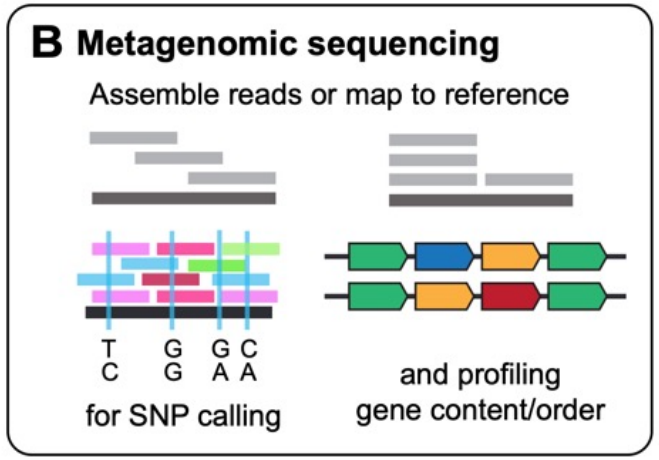








How to study strains in a culture-free context?

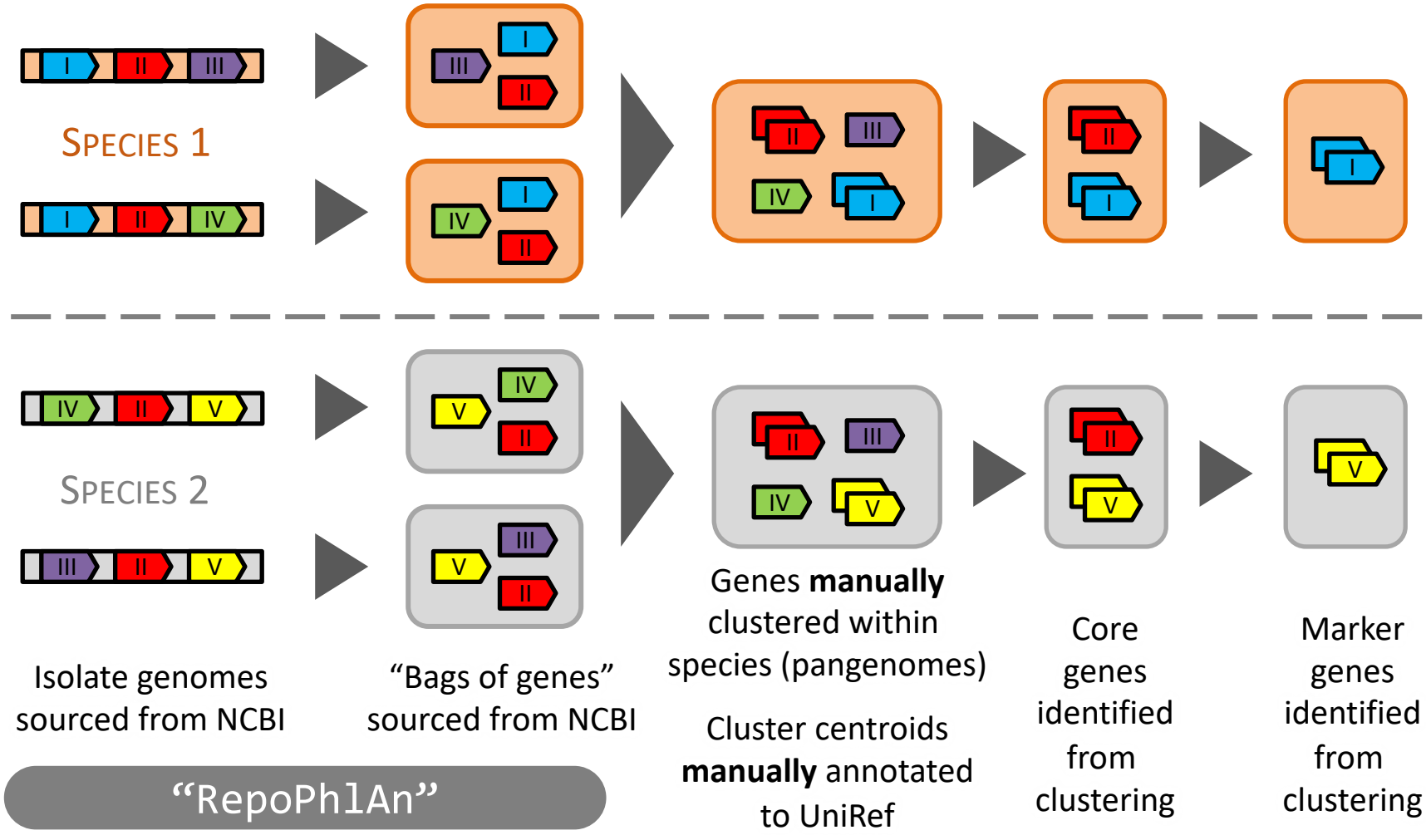


Reference-based strain identification

- Four broad mechanisms for reference level strain profiling
  - Identification of reference genotypes in a community (e.g. PathoScope; Sigma)
  - Identification of the dominant strain (incl. novel) per species (e.g. StrainPhlAn, MetaMLST, MetaSNV)
  - Identification of more than one strain per species (e.g. ConStrains, DESMAN, StrainGR)
  - Identification of structural variants within a community (e.g. PanPhlAn)

Gene-level strain profiling with PanPhlAn

Pangenomes and marker genes as strain profiling tools



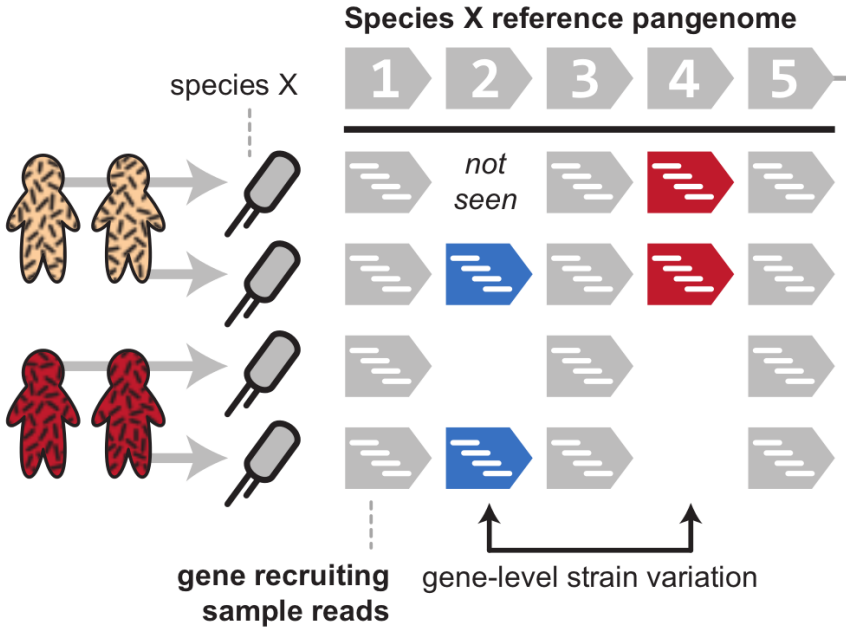
- RepoPhlAn and ChocoPhlAn are database systems for identifying markers genes from isolate genomes (among other tasks).
- We will use other products of these systems in subsequent analysis methods.

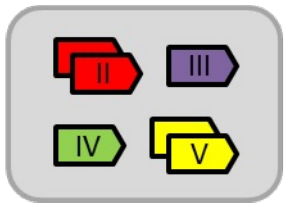
Stable, individualized marker loss hints at strains

Franzosa, PNAS, 2015

Gene-level strain profiling

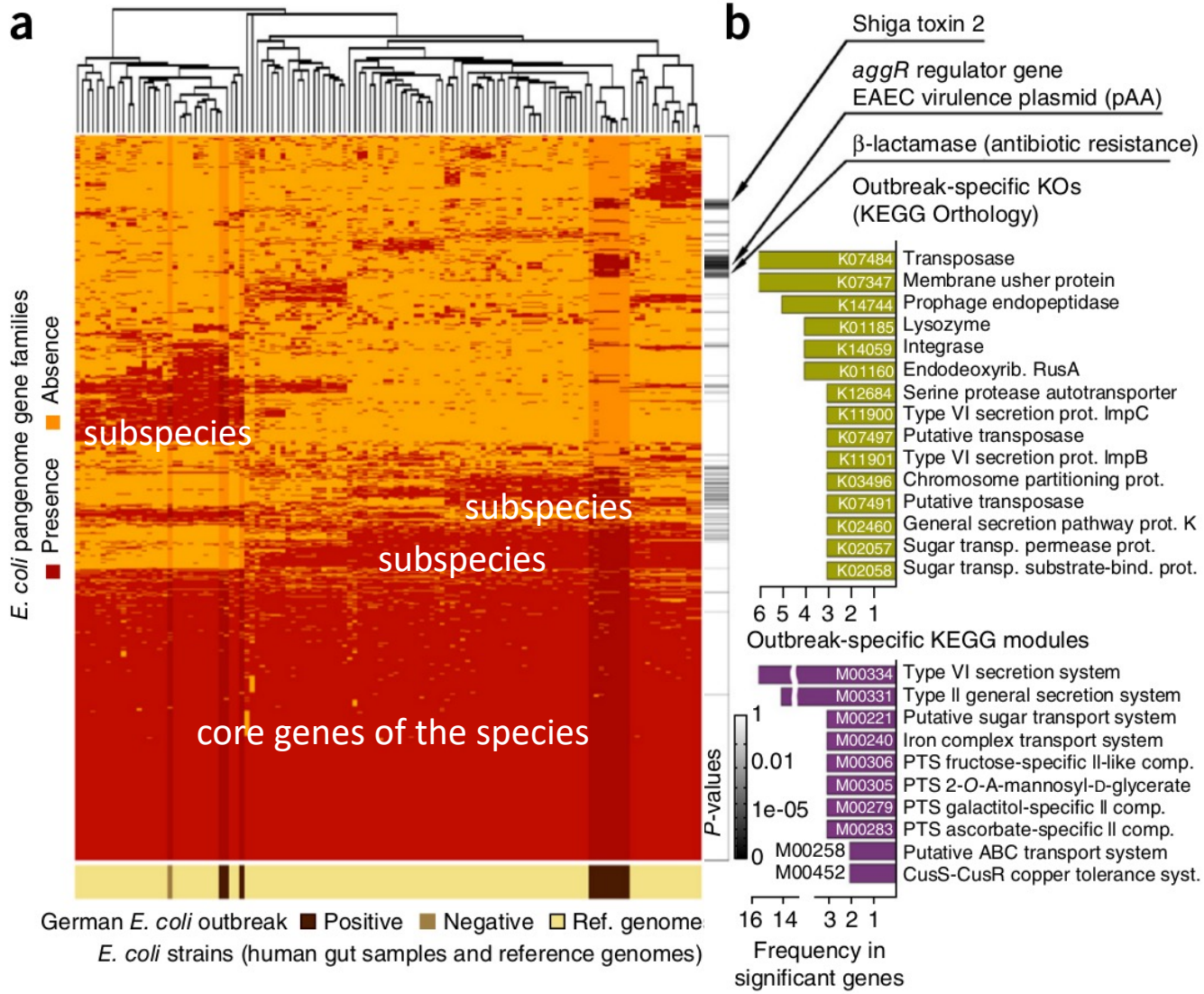
Associate strain-variable genes with phenotypes/functions



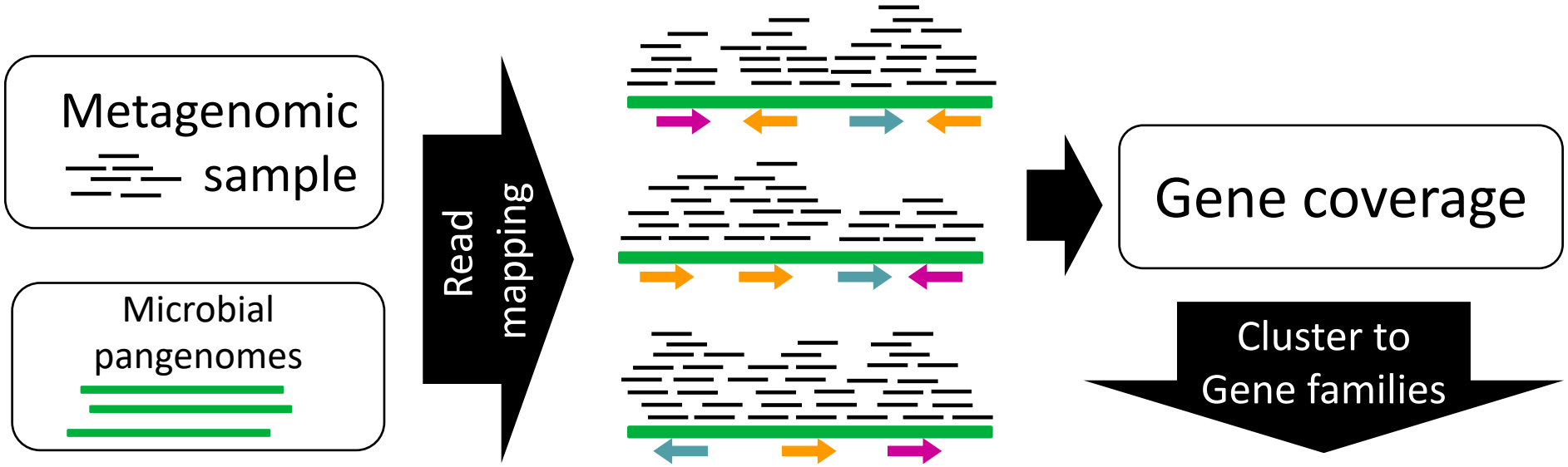


Clustered pangenomes with gene families

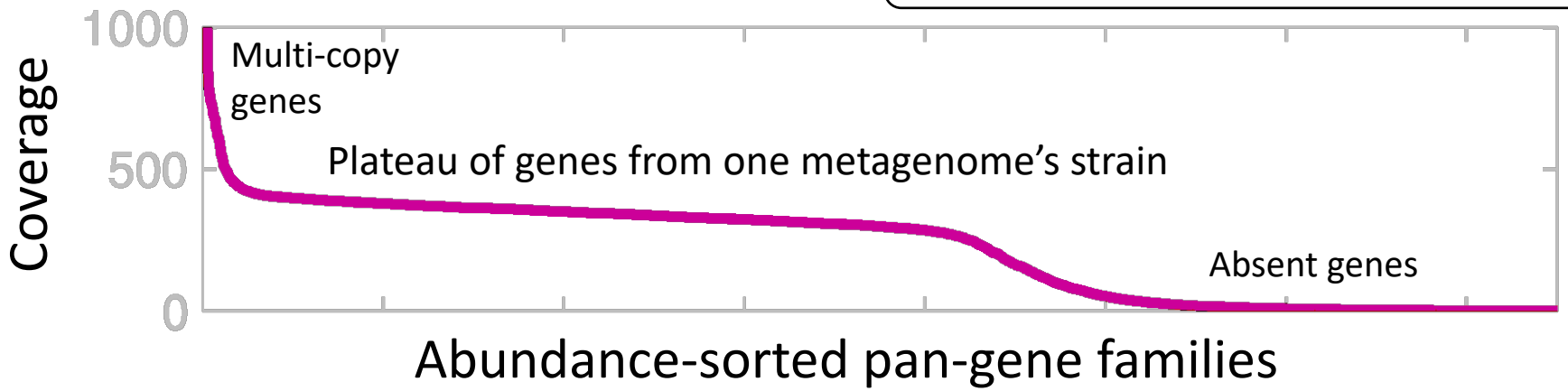
Look for variation in gene presence across samples after mapping reads to species' pangenomes



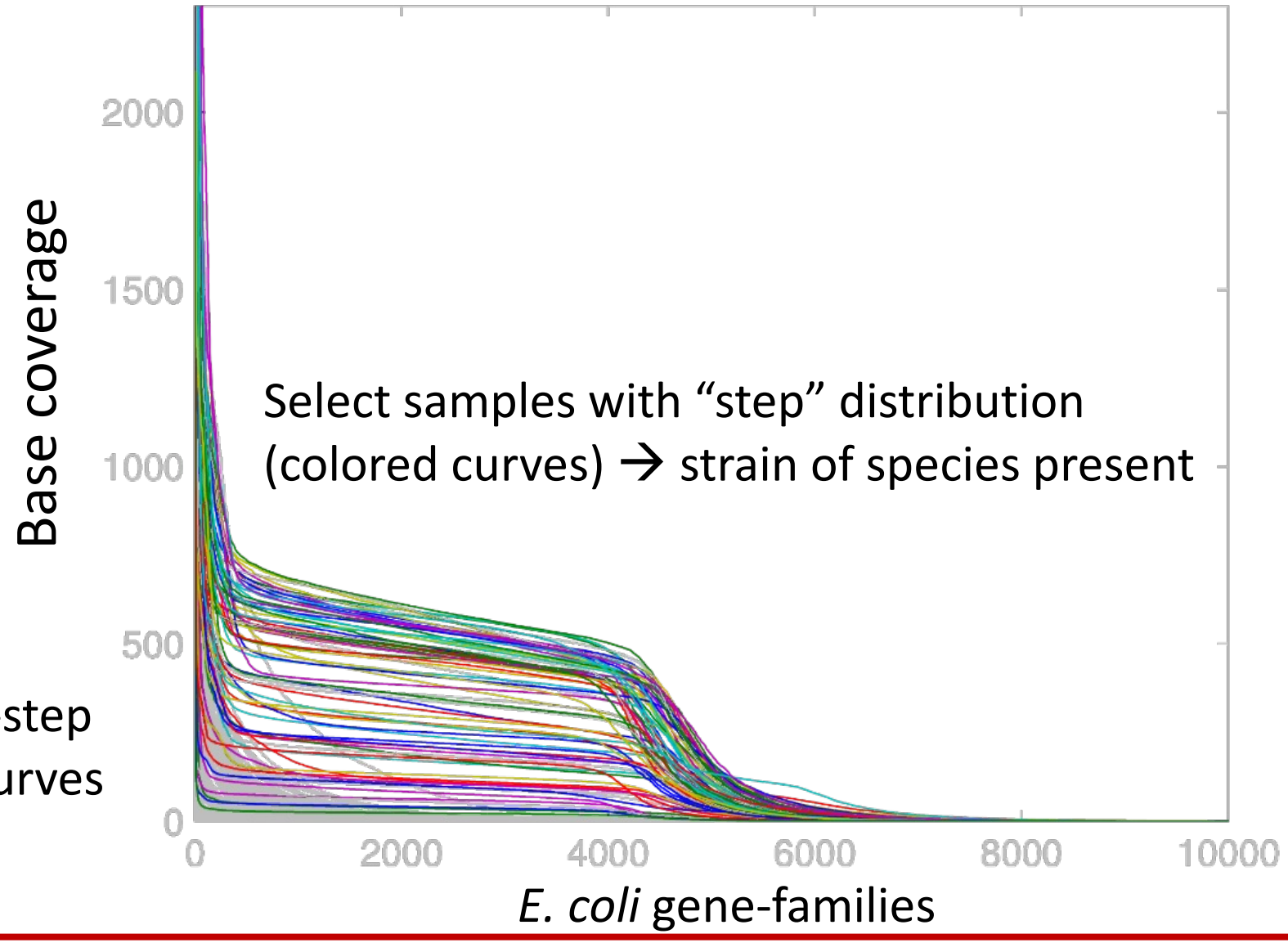
The PanPhlAn approach



Pan-gene family coverage



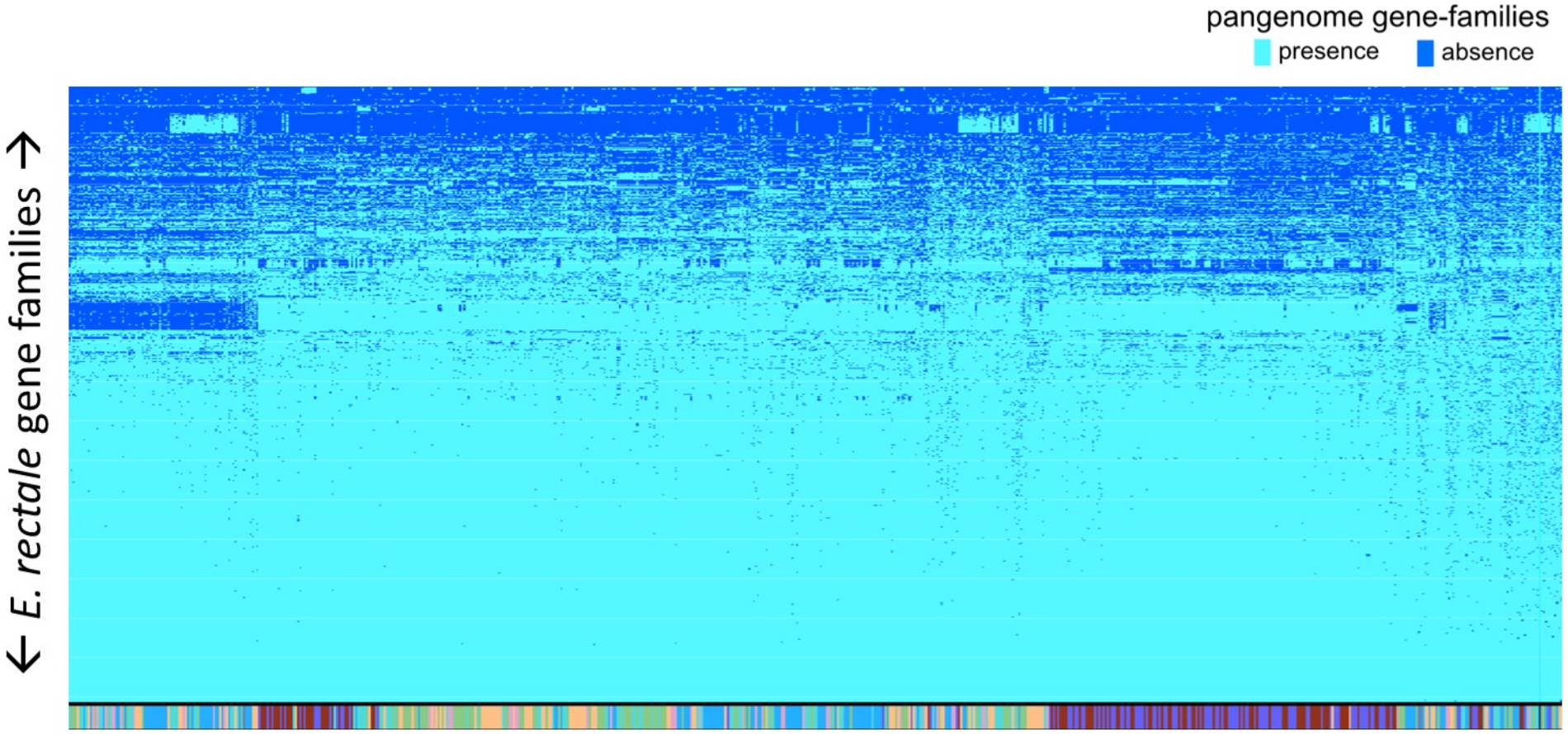
The PanPhlAn approach



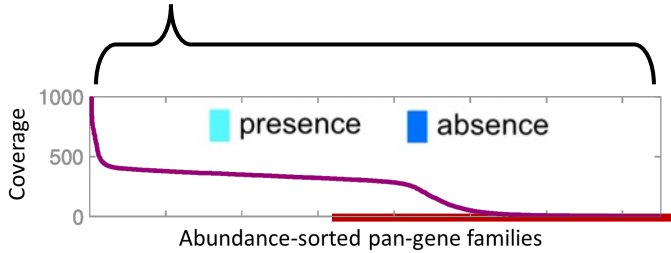
Select samples with “step” distribution (colored curves) → strain of species present

Reject non-step (gray) curves

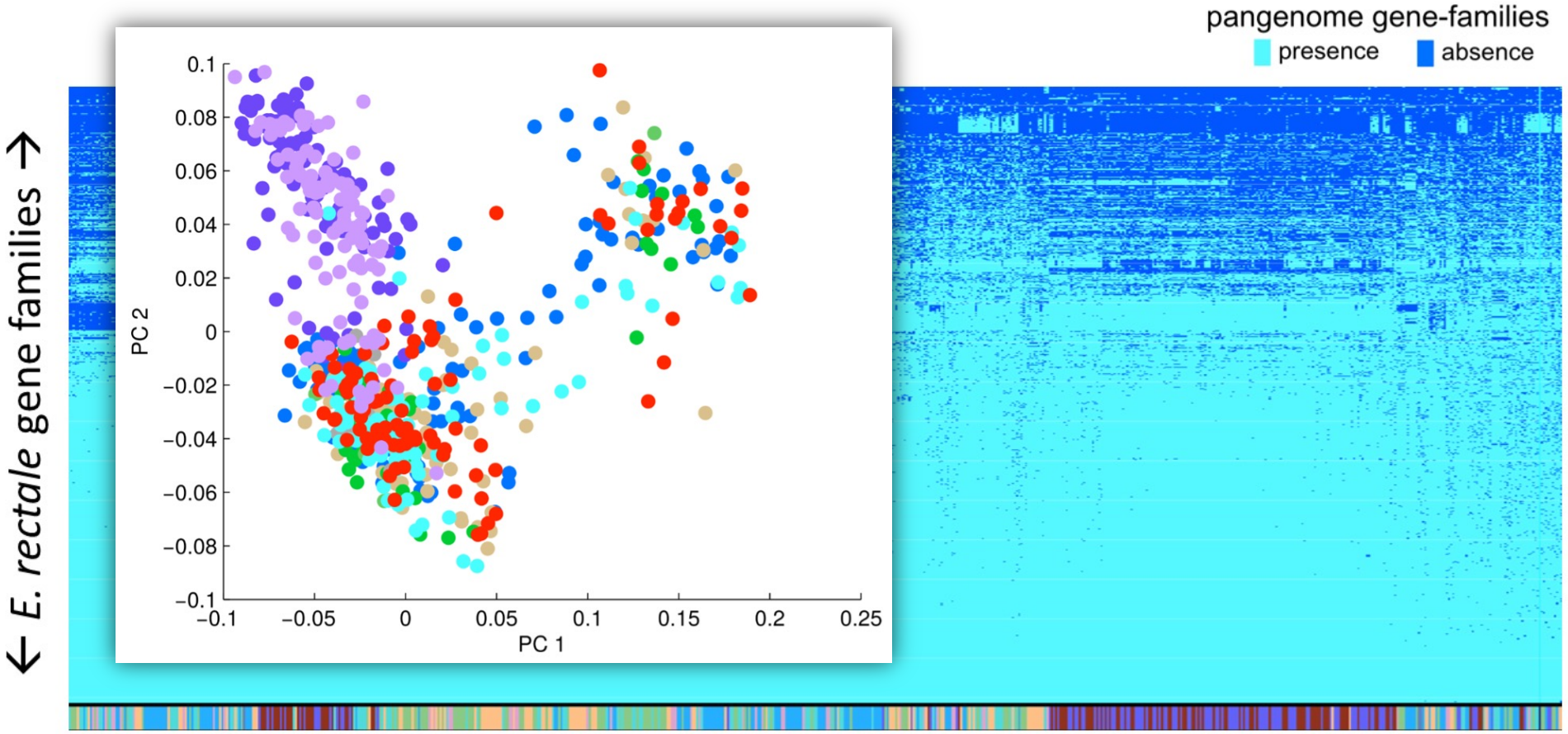
PanPhlAn reveals strain-level population structure of *Eubacterium rectale* from a single reference genome



Each skinny column represents a plateau of one species in one sample binarized to gene presence and absence



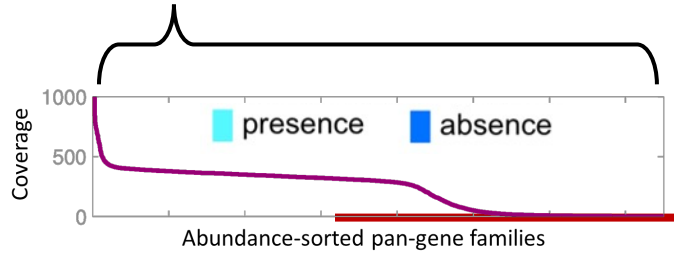
PanPhlAn reveals strain-level population structure of *Eubacterium rectale* from a single reference genome



Each skinny column represents a plateau of one species in one sample binarized to gene presence and absence

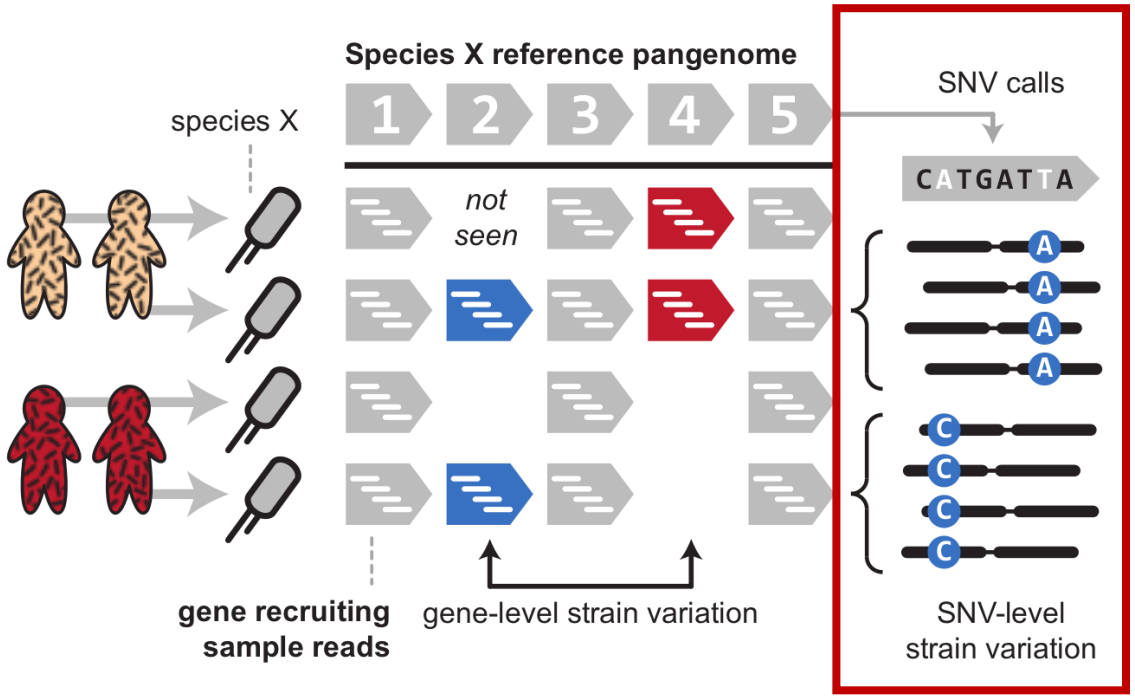
← 1000s of human gut metagenomes →

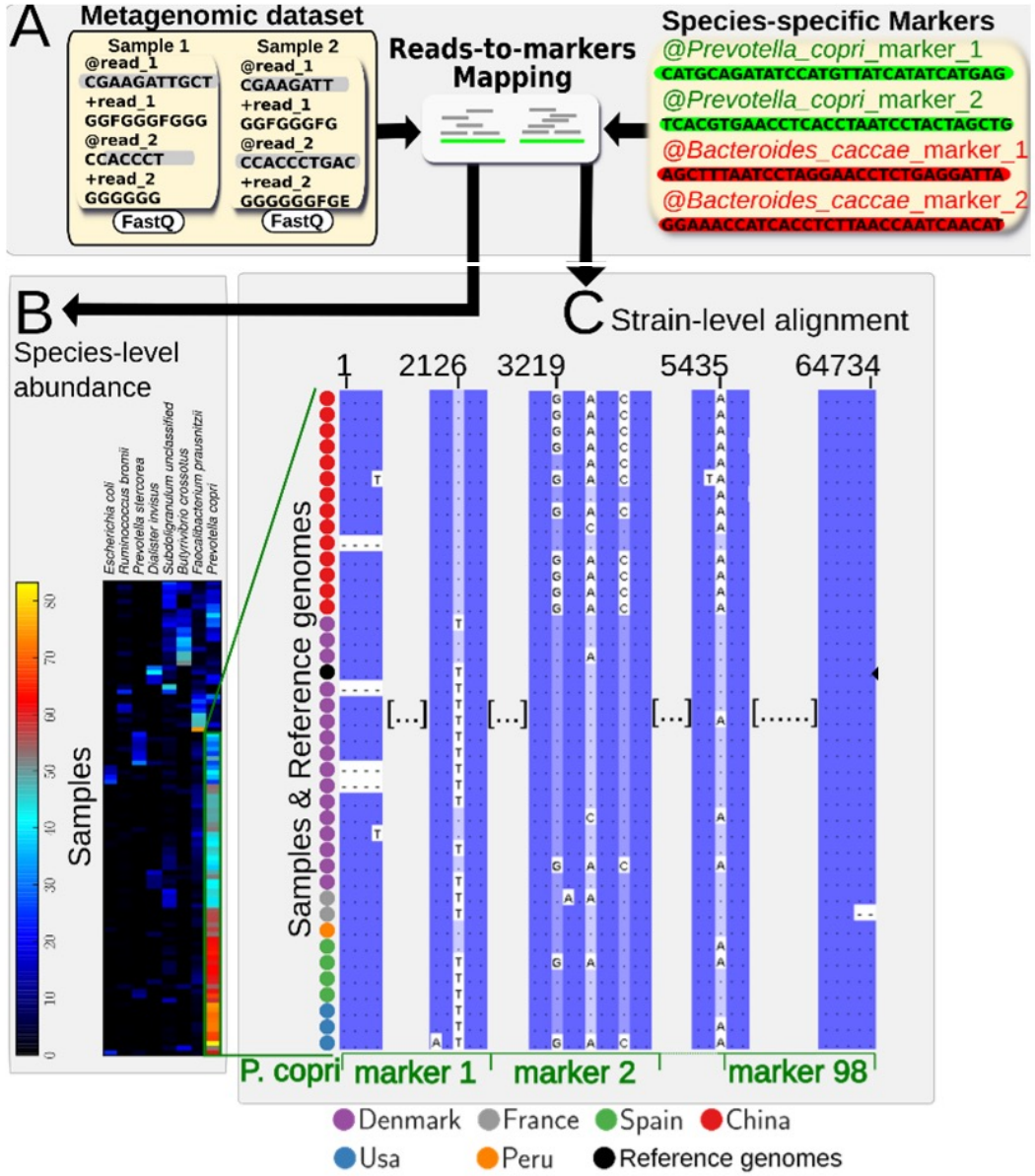
HMP
Chinese T2D
WT2D
Neilsen/metahit
E. coli outbreak
Infants
Zeller cancer
Chatelier
Chinese liver



SNV-level strain profiling with StrainPhlAn

Profiling strains at nucleotide-level resolution





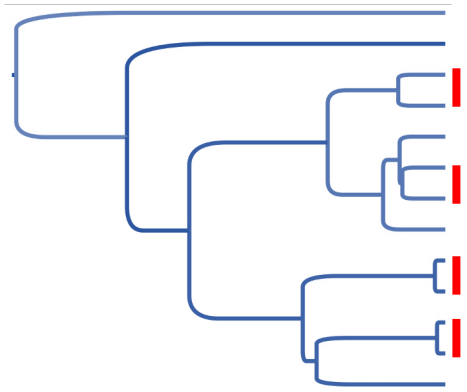
- **StrainPhlAn** reconstructs species' dominant per-sample haplotype from SNVs in species-specific marker genes
- For each species, build a strain distance matrix over samples

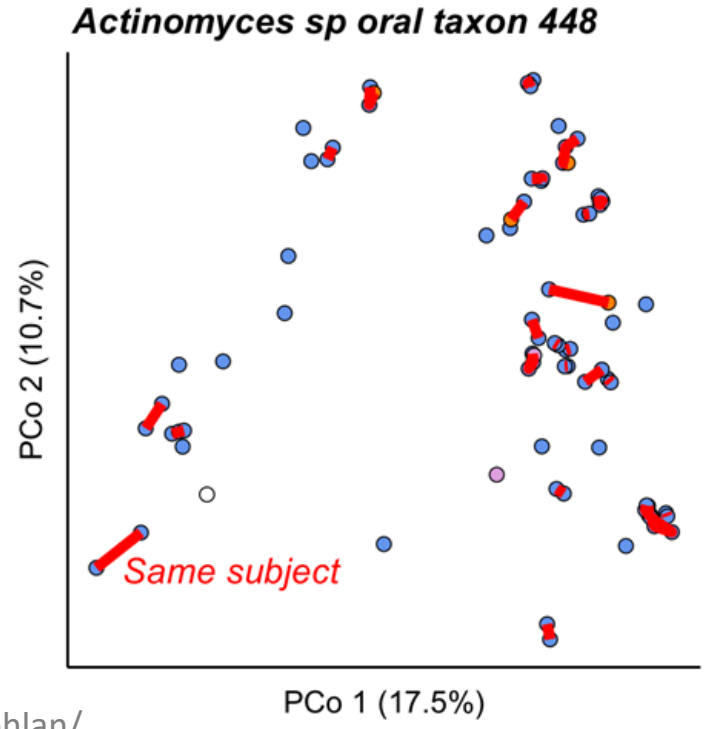
	Sample j
Sample i	$d(i,j)$

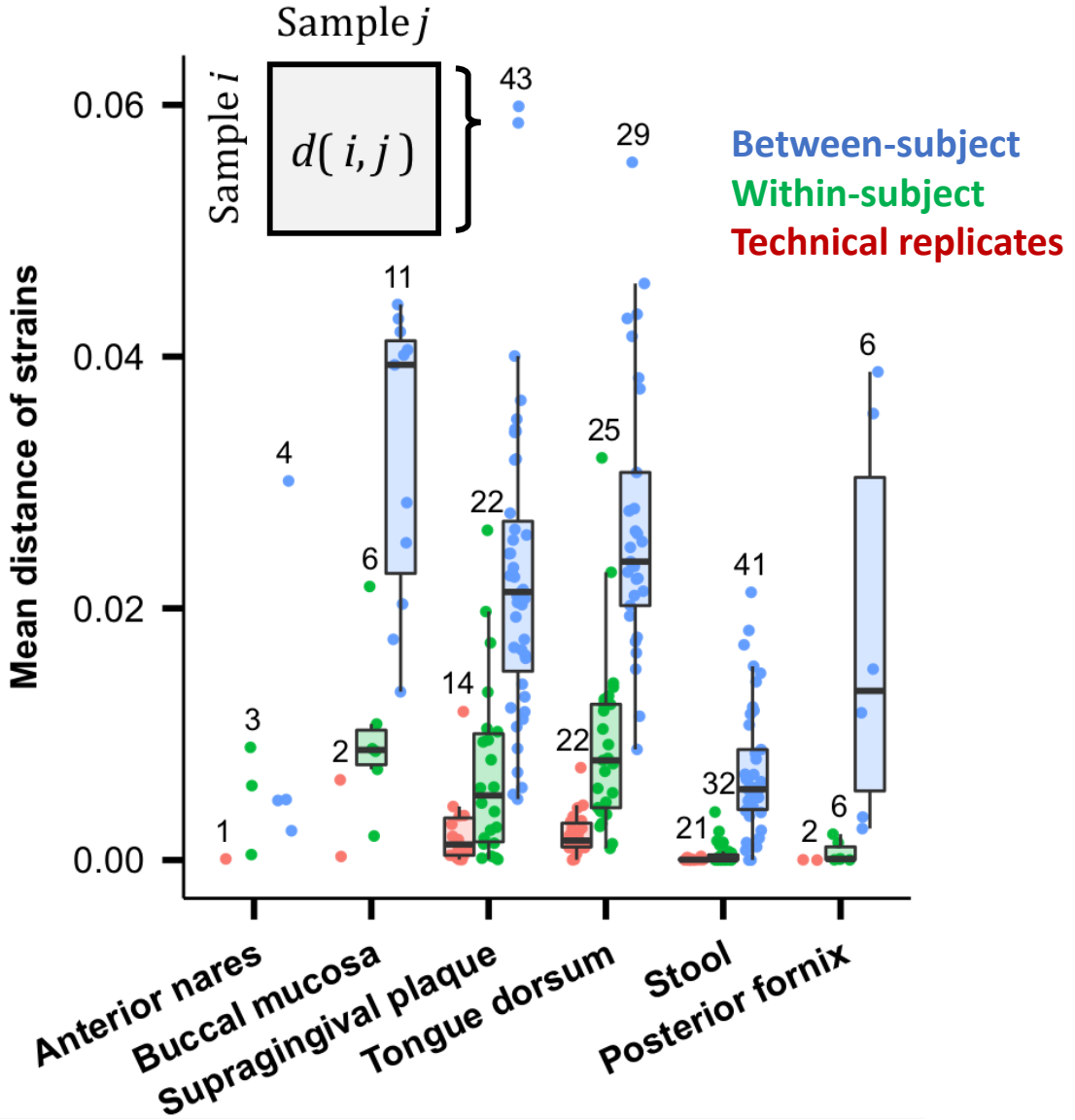
Strain-level properties of the human microbiome

- Less haplotype diversity among stool species compared to oral sites
- *d*(between-subject)

  >> *d*(within-subject) > *d*(tech. reps.)



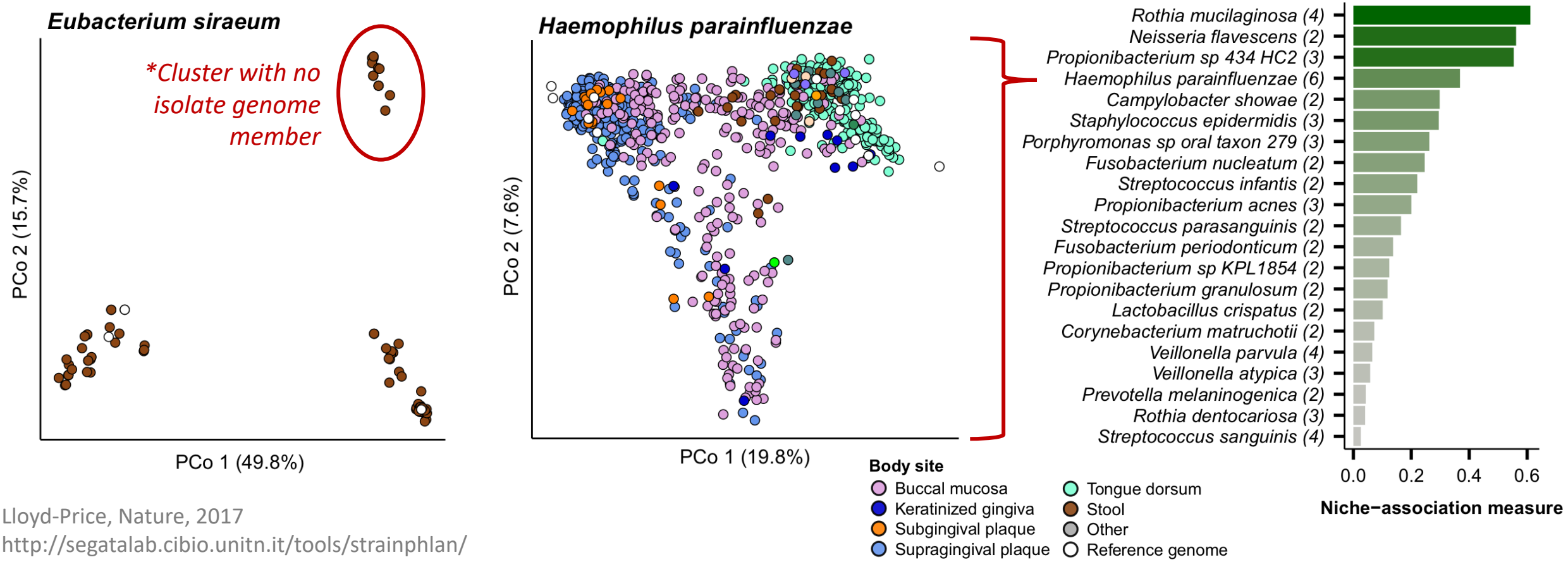




Lloyd-Price, Nature, 2017
http://segatalab.cibio.unitn.it/tools/strainphlan/

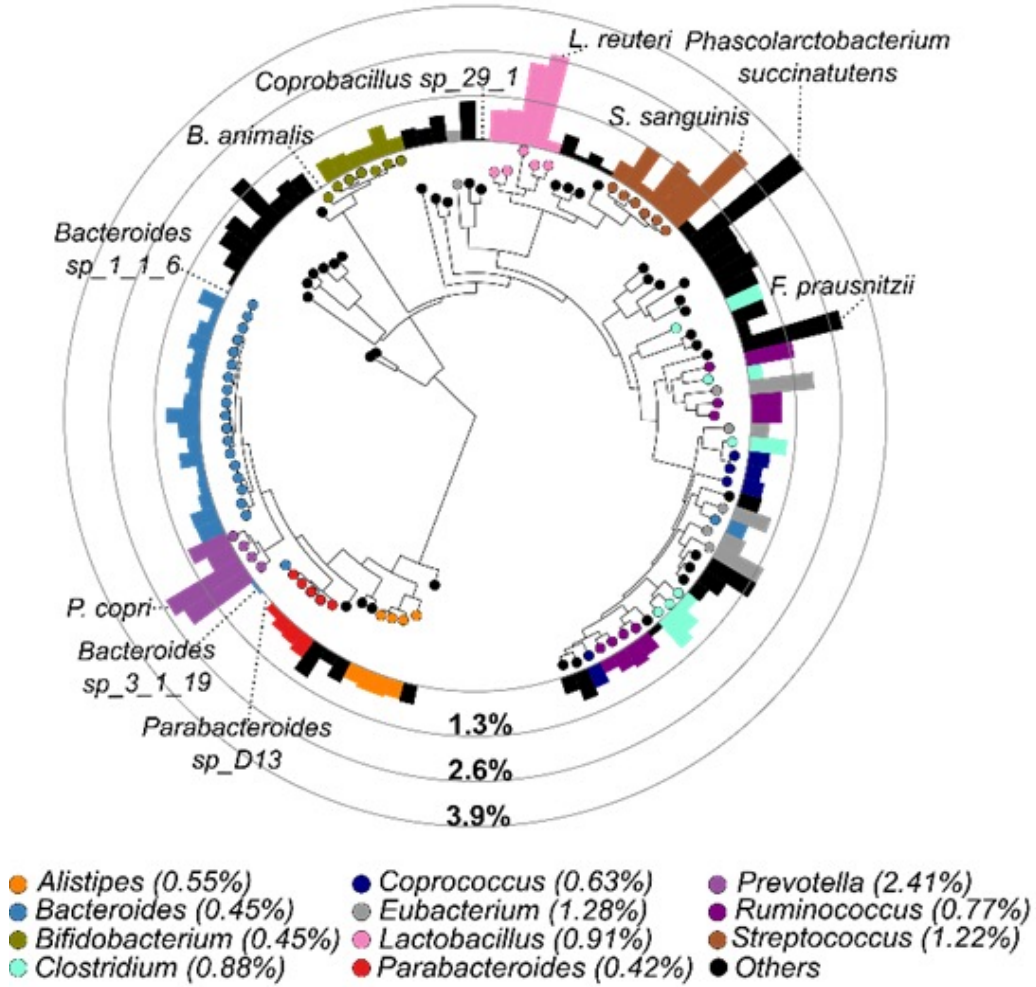
Strain-level properties of the human microbiome

- Subspecies structure manifests as “clusters” of strains
- Structure associated with body site suggests possible niche-adaptation

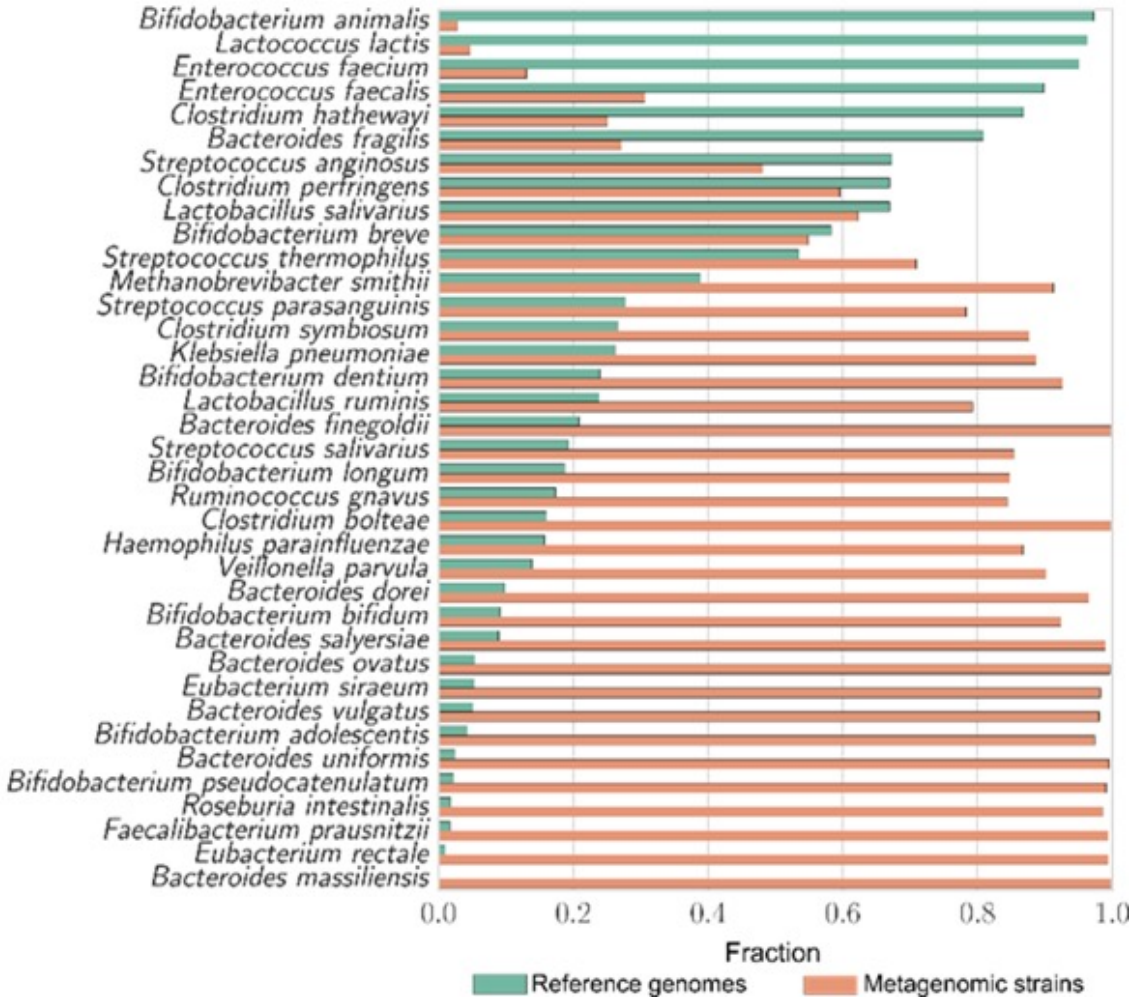


Lloyd-Price, Nature, 2017
http://segatalab.cibio.unitn.it/tools/strainphlan/

Lots of undersampled species genome diversity



Median divergence from reference markers

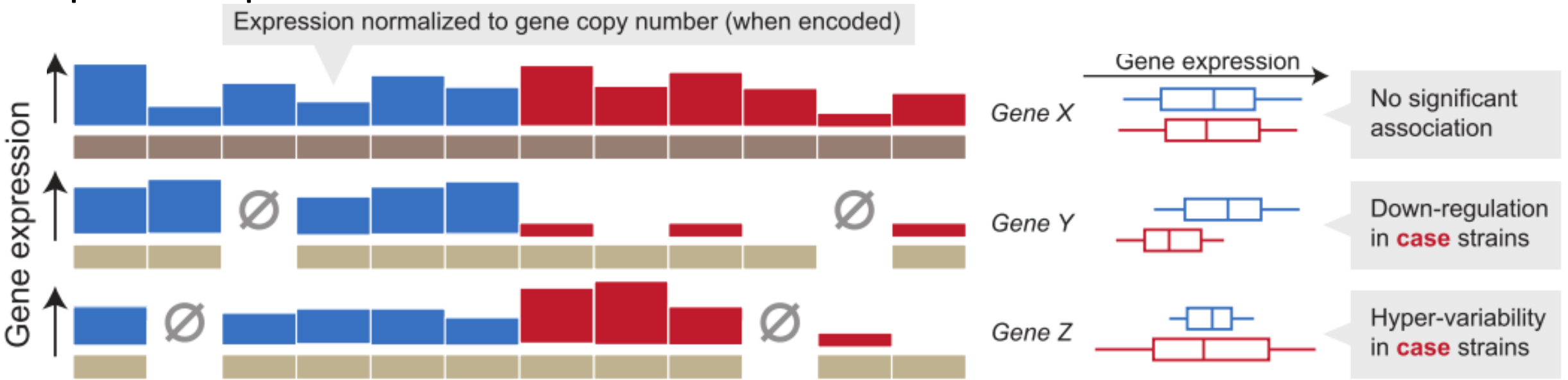


Fractions of intraspecific branch length for genomes/metagenomic strains

Metatranscriptomics

Strain tracking by gene expression

- Strain identification by analogous differences in metatranscriptomic gene expression profiles



- Limited usage to date one of the biggest findings has been on the ability of certain *E.lenta* strains to metabolize digoxin (Haiser et al, 2014 Science)
- Use has expanded in the last couple of years for viral identifications and novel strain finding

Sardi et al, 2020 Viruses

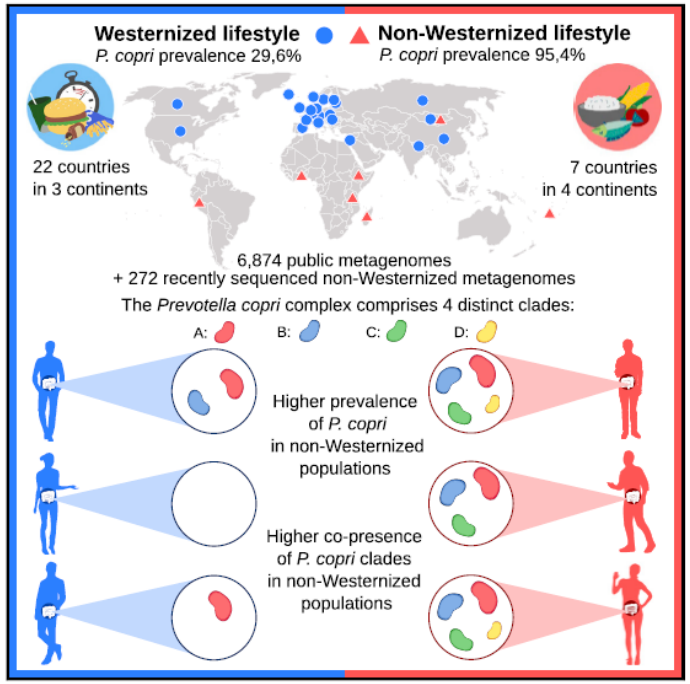
A world of strain biology within individual species

Article

Cell Host & Microbe

The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations

Graphical Abstract



Authors

Adrian Tett, Kun D. Huang,
Francesco Asnicar, ...,
Curtis Huttenhower, Frank Maixner,
Nicola Segata

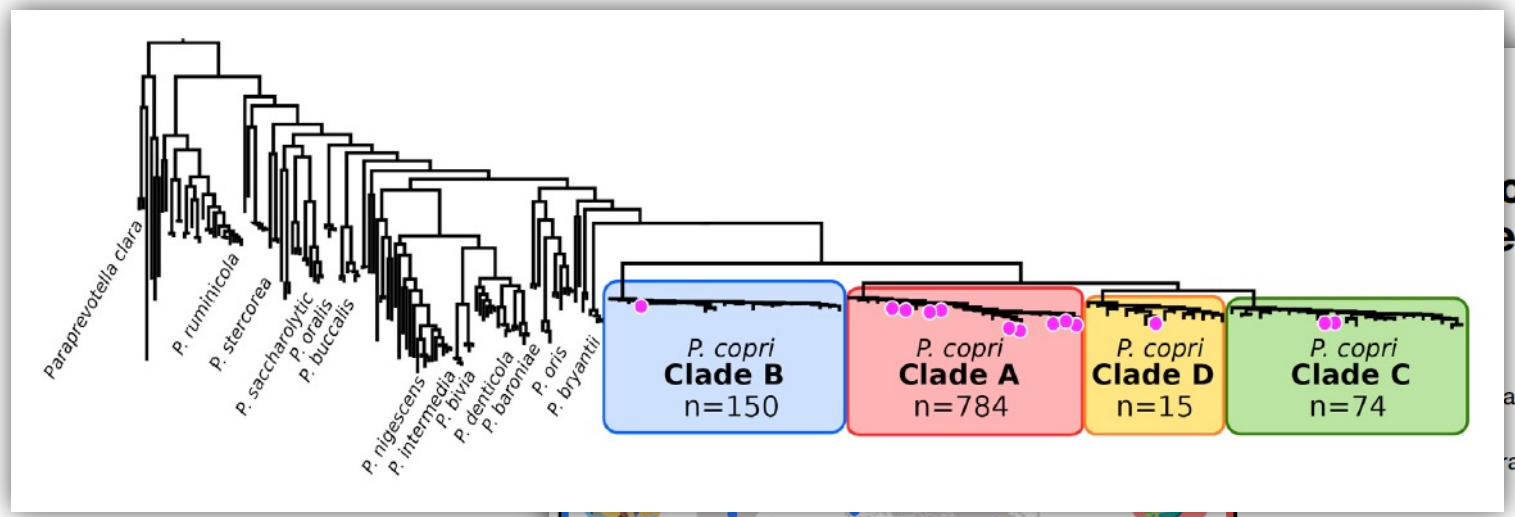
Correspondence

adrianjames.tett@unitn.it (A.T.),
nicola.segata@unitn.it (N.S.)

In Brief

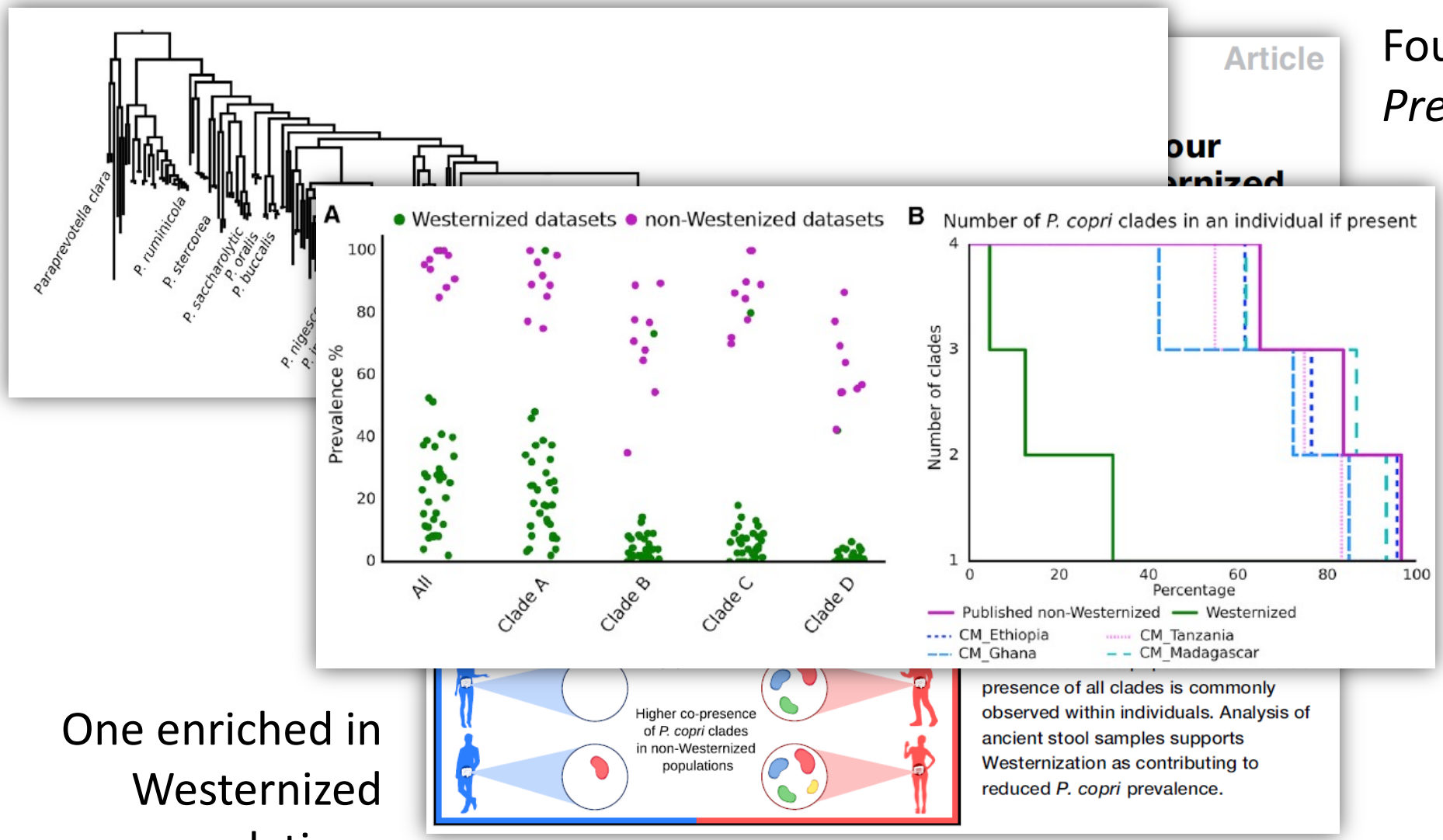
Tett et al. find that the intestinal microbe *Prevotella copri* encompasses four distinct clades constituting the *P. copri* complex. The complex is prevalent in non-Westernized populations where co-presence of all clades is commonly observed within individuals. Analysis of ancient stool samples supports Westernization as contributing to reduced *P. copri* prevalence.

A world of strain biology within individual species



Four clades within the *Prevotella copri* species

A world of strain biology within individual species

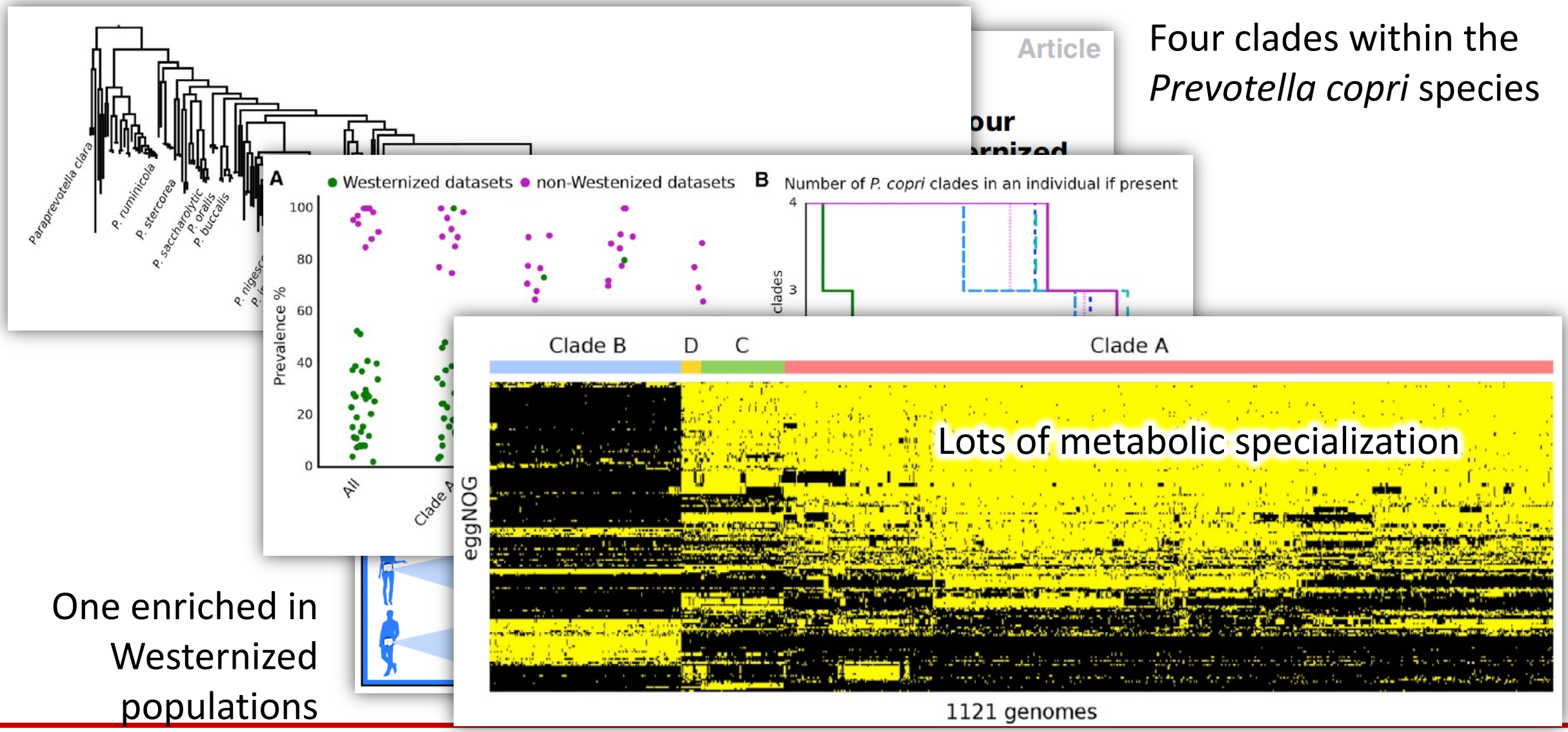


Four clades within the *Prevotella copri* species

One enriched in Westernized populations

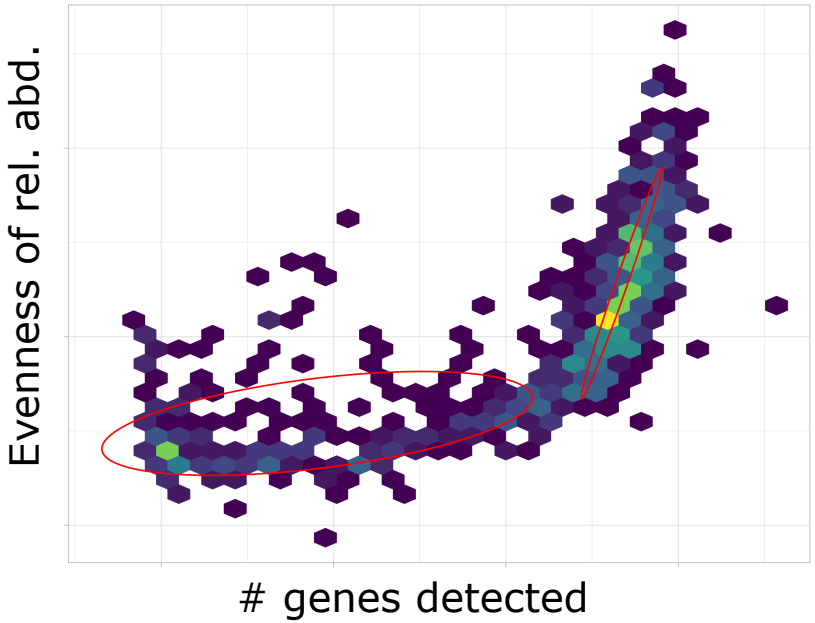
A world of strain biology within individual species

Four clades within the *Prevotella copri* species

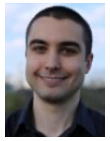


One enriched in Westernized populations

ANPAN: METHODS FOR STRAIN-LEVEL BIOMARKER DISSECTION



ANPAN

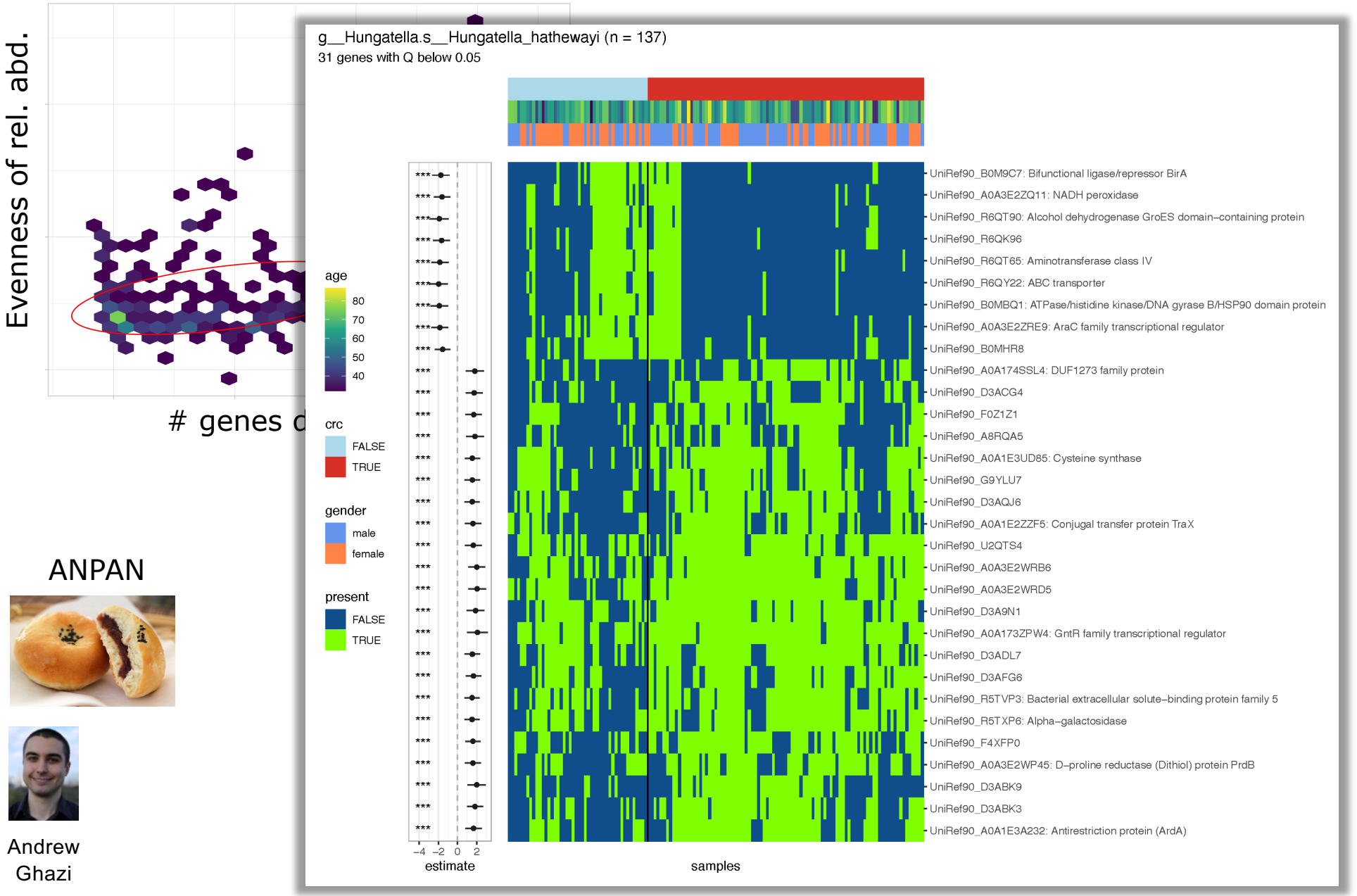


Andrew Ghazi

ANPAN: METHODS FOR STRAIN-LEVEL BIOMARKER DISSECTION

CANCER GRAND CHALLENGES

Team: OPTIMISTICC



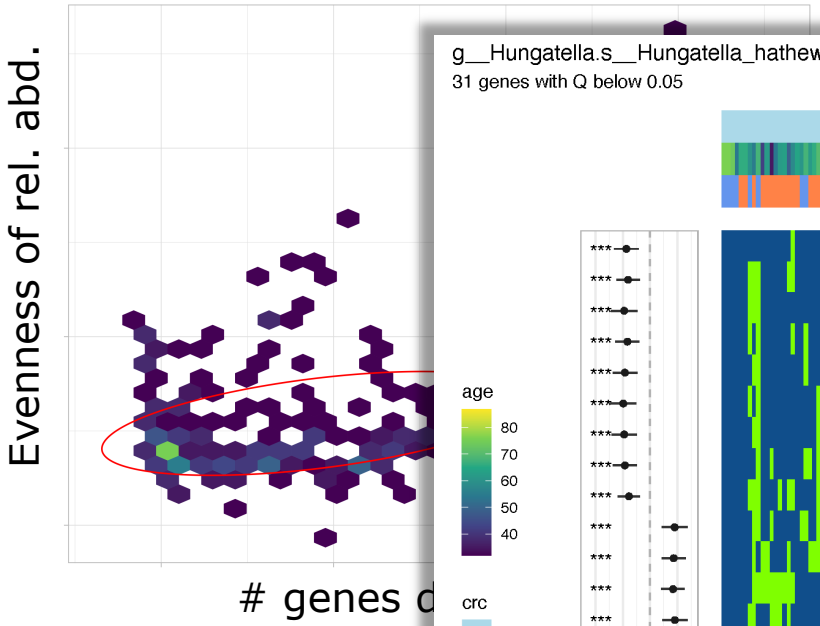
ANPAN

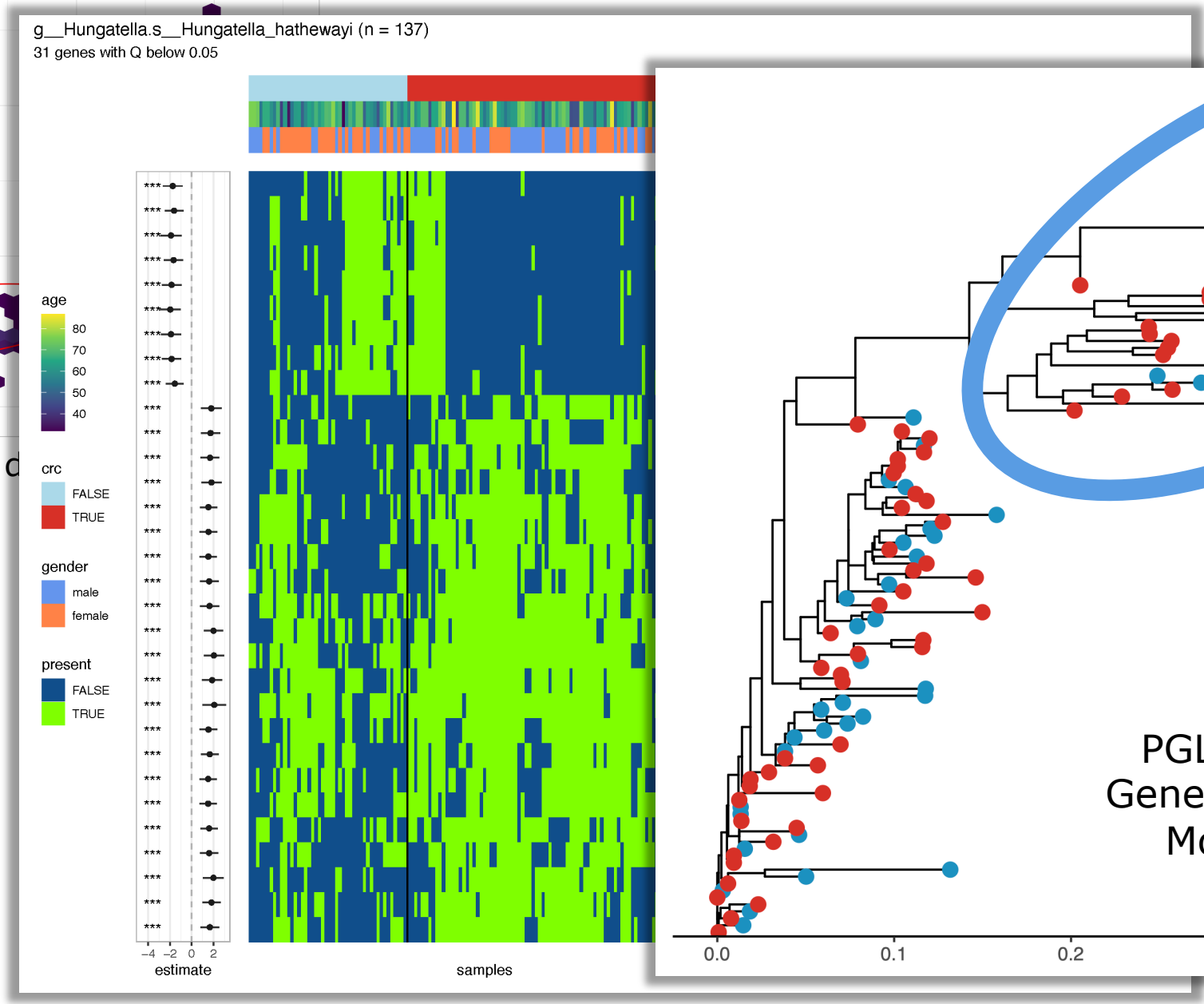
Andrew Ghazi

ANPAN: METHODS FOR STRAIN-LEVEL BIOMARKER DISSECTION

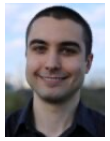
CANCER GRAND CHALLENGES

Team: OPTIMISTICC





ANPAN



Andrew Ghazi

Assembly-based approaches

Haha fooled you – that’s tomorrow

Summary

- What is a microbial strain?
- Why is considering strains important?
- Gene-level strain profiling with PanPhlAn
- SNV-level strain profiling with StrainPhlAn
- Community strain statistics with ANPAN

The bioBakery team(s)

The Laboratory of Computational Metagenomics

University of Trento, Italy

http://segatalab.cibio.unitn.it

The Huttenhower Lab

Harvard Chan School of Public Health

http://huttenhower.sph.harvard.edu